The Development of Responses to Unfairness in Children

by

Young-eun Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psychology)
in the University of Michigan
2021

Doctoral Committee:

Associate Professor Felix Warneken, Chair
Assistant Professor Julia L. Cunningham
Professor Susan A. Gelman
Professor Henry M. Wellman

Young-eun Lee

leeyeun@umich.edu

ORCID iD: 0000-0002-1392-0514

Dedication

To my family who has been unconditionally supporting me.

Acknowledgements

I would like to thank my advisor Felix Warneken for his dedication and commitment to providing me excellent mentorship and advice. This dissertation would not exist without his support, guidance and patience. Thank you for helping me grow as a researcher and as a person. I am grateful to the members of my committee, Susan Gelman, Henry Wellman and Julia Lee Cunningham, for their encouragement, feedback and mentorship. Also, this dissertation would not have been possible without help from research assistants, child participants and their families. I would like to thank my friends, past and present lab mates who have helped me along the way. Lastly, I thank my family for their unwavering support.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables v

List of Figures vi

Abstract vii

Chapter 1: Introduction 1

Chapter 2: Study 1 9

Chapter 3: Study 2 24

Chapter 4: Study 3 54

Chapter 5: Study 4 66

Chapter 6: General Discussion 103

Bibliography 107

List of Tables

Table 1. Experimental design with resource allocation (Divider:Recipient) 72

Table 2. Mean proportion of children who chose helpers over punishers (Study 4A) and those who chose rational over irrational third parties (Study 4B) in binomial tests collapsed across age. 79

List of Figures

Figure 1. Schematic representation of the proposed research. 7

Figure 2. Computer screen display of the coin game during the test phase in Study 1. 13

Figure 3. Children's punishment decisions in Study 1. 17

Figure 4. Number of coins taken away from players in Study 1. 20

Figure 5. Schematic representation of the design in Study 2A. 31

Figure 6. Children's punishment decisions in Study 2A. 37

Figure 7. Schematic representation of the design in Study 2B. 41

Figure 8. Children's expectation about a divider's offer before and after the second-party game. 45

Figure 9. Children's punishment decisions in Study 2B. 47

Figure 10. Computer screen display of the coin game in the same divider condition. 58

Figure 11. Computer screen display of the coin game in the different divider condition. 59

Figure 12. Children's punishment decisions in Study 3. 63

Figure 13. Schematic set-up of stimuli used in Study 4A. 71

Figure 14. Likert scale ratings and social preference scores in Study 4A and 4B. 78

Figure 15. Likert scale ratings in Study 4C and 4D. 90

Figure 16. Forced-choice social preferences in Study 4D. 96

Abstract

In our daily life, we often experience moral outrage when we hear news about perpetrators who treated others unfairly even if we are not the victim. We also think that perpetrators should receive punishment they deserve. In fact, research shows that adults are often willing to pay a cost to intervene against such fairness norm violations even when they are an uninvolved third party. This so-called third-party punishment is striking because it cannot be easily explained by self-interested motivations. If people were rational agents who try to maximize their own payoffs, they would not pay any costs to intervene in third-party transgressions. Thus, third-party punishment has often been considered as an index of one's sense of fairness. However, despite its theoretical importance, in the field of developmental psychology, its underlying mechanisms and developmental trajectories have been relatively understudied. This dissertation includes four sets of studies to assess following questions: (a) When do children start to engage in and reason about third-party punishment? (b) What motivates third-party punishment in children?

To answer the first question, by testing a wide age range (age 5 to 9), I found that with age, children's punishment becomes increasingly selective (Study 1 & 2). That is, over development, children are less likely to punish fair allocations, while they become more likely to punish unfair allocations. Further, from age 7, children start to think of third-party punishment as a way to reduce inequality between two other individuals (Study 4). To answer the second question, I examined the influences of children's own experience (Study 2) and the possibility of

future interactions (Study 3) on third-party punishment, respectively. I found that neither robustly influenced third-party punishment in children. Rather, children enact third-party punishment in a way that could restore equality between two other people (Study 1), suggesting that their punishment is motivated by fairness concerns. However, despite children's use of punishment to rectify inequality, I found that children prefer third-party helpers over third-party punishers (Study 4), which questions the extent to which children endorse third-party punishment as an appropriate intervention against unfairness.

Taken together, four sets of studies suggest that third-party punishment in children reflects their fairness concern. This dissertation elucidates the development and motivations of third-party punishment in children.

Chapter 1: Introduction

Cooperation is an important feature of human social groups. While humans show extraordinary abilities to cooperate, there is often a tension between an individual's interests and those of other social partners. Fairness norms serve as a guide on how to resolve this tension, providing standards of behavior that can foster cooperation and help prevent individuals from undermining social relationships. However, the mere existence of fairness norms is not sufficient, as individuals might follow them imperfectly or not at all. Therefore, both theoretical models and empirical research highlight how different forms of intervention against those who violate fairness norms are important. Specifically, in direct interactions, individuals might retaliate against those who treat them unfairly or shun them to avoid further exploitation (Balliet, Mulder, & Van Lange, 2011; Baumard, André, & Sperber, 2013; Fehr & Fischbacher, 2004a). Moreover, in group contexts, individuals might punish free-riders who contribute less to a public good than others, a mechanism that maintains a higher level of cooperation in the face of defection (Fehr & Gächter, 2002; Gurerk, Irlenbusch, & Rockenbach, 2006). These findings suggest that people enact punishment when cooperative norms are violated.

Importantly, people punish those who violate fairness norms not only when their interest is at stake but also when they are an uninvolved observer. For example, studies show that adults punish unfair resource dividers even in situations where they are a third-party who is not directly affected by the unfair allocations (Fehr & Fischbacher, 2004a; Henrich et al., 2006; Krasnow, Delton, Cosmides, & Tooby, 2016; Nelissen & Zeelenberg, 2009; Yamagishi et al., 2017). More

strikingly, adults are willing to pay a personal cost in such third-party situations by e.g. paying their own money to inflict costs on a perpetrator. This phenomenon is called *costly third-party punishment* (TPP) and is well-established in adults.

The underlying motivations for this so-called third-party punishment remains hotly debated. One theory suggests that punishment of unfair sharing is an index of one's altruistic tendencies and concern for group norms (e.g., Fehr, Fischbacher, & Gächter, 2002; Fehr & Fischbacher, 2003; Fehr & Fischbacher, 2004b). Whereas, other theorists suggest that the underlying motivation for punishment is ultimately self-interested (Krasnow et al., 2016; Petersen, Sell, Tooby, & Cosmides, 2010) or spiteful (Yamagishi et al., 2012, 2017). Importantly, despite disagreements over what exactly are the underlying motives, costly punishment has been identified as an important phenomenon in the study of human cooperation because it addresses the issue of how to respond to acts of free-riding (Balliet et al., 2011; Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Gächter, 2002; Gurerk, Irlenbusch, & Rockenbach, 2006). In fact, TPP has often been claimed to reflect one's concern for fairness norms because a third-party pays a personal cost to punish the perpetrator with no immediate benefits (Fehr & Fischbacher, 2003; Fehr & Fischbacher, 2004b, but see Raihani & Bshary, 2019 for a review of a different view).

Empirical Evidence About Third-Party Punishment in Children

TPP marks a developmental milestone in fairness development in that children have to overcome their self-interest and apply the fairness norms even when their interest is not at stake (McAuliffe, Blake, Steinbeis, & Warneken, 2017). As young children tend to have a self-serving bias in resource allocations, it is important to understand when and how children's genuine concern for fairness norms emerges. For example, in one study (Smith, Blake, & Harris, 2013),

children as young as 3 years old report that they should share resources equally with others, acknowledging the fairness norms. However, it is not until age 7 to 8 that children share resources equally, suggesting a discrepancy between their understanding of fairness norms and behavioral adherence to these norms. Similarly, 8-year-olds, but not 4- to 7-year-olds, avoid not only receiving fewer resources than a partner but also receiving more resources than the partner (Blake & McAuliffe, 2011), implying that it is not until age 8 that children apply fairness norms when their payoff is at an advantage. Therefore, TPP could be a critical test case for the emergence of fairness concerns both across phylogeny and ontogeny because children have to incur a personal cost to punish unfair allocations when there are no immediate benefits to the self (McAuliffe et al., 2017).

Developmental studies have begun to trace the developmental trajectory of TPP in children. Two studies using looking-time measures have found that infants expect a differential treatment of prosocial and antisocial agents. For example, 10-month-olds look longer at an event in which an unfair resource divider receives punishment than an event in which a fair resource divider receives punishment (Meristo & Surian, 2014) and 13- to 15-month-old infants differentially associate verbal praise and admonishment with fair and unfair resource dividers (Deschamps, Eason, & Sommerville, 2015). While these studies explore infants' event representations and how they anticipate others would act towards fair or unfair behavior, they do not speak to infants' evaluations of these behaviors or their own third-party intervention.

Several studies have shown that a precursor of TPP can be found in children aged 2 to 3 years of age. For example, 2-year-olds take treats away more often from an agent who previously hindered another agent than from a helpful agent (Hamlin, Wynn, Bloom, & Mahajan, 2011). Furthermore, 3-year-olds protest verbally against a person who destroyed another person's

belonging (Vaish, Missana & Tomasello, 2011). Children around this age also punish the destroyer by taking his or her opportunity to engage in a fun, desirable activity (Yudkin, Van Bavel, & Rhodes, 2020). Overall, these studies reveal that by 3 years of age, children actively intervene against moral transgressions in some contexts, even if they are an unaffected third-party observer. While these studies examine interventions against certain forms of antisocial behavior, such as hindering of another agent's instrumental goal (Hamlin et al., 2011) or damage to one's property (Vaish et al., 2011; Yudkin et al., 2020), these prior studies did not assess how children react to a situation in which fairness norms are violated.

Children's TPP against unfairness has been found in children aged 6 and older (Gummerum & Chu, 2014 with samples from the UK; House et al., 2020 with samples in Argentina, India, Germany and the US; Jordan, McAuliffe, & Warneken, 2014; McAuliffe, Jordan, & Warneken, 2015 with samples in the US). Studies have shown that when in the role of a third party, 6- to 8-year-olds are willing to pay a cost to prevent inequality (Gummerum & Chu, 2014; Jordan, McAuliffe, & Warneken, 2014; McAuliffe, Jordan, & Warneken, 2015). Specifically, in McAuliffe et al. (2015), children intervened selectively when dividers made unequal, selfish offers to a recipient (and did not intervene when the divider split resources equally with the recipient). This study found that this pattern emerges robustly by 6 years of age, with 5-year-olds already trending in the same direction but not yet reliably punishing unfairness.

The studies illustrated above measured children's costly punishment as a binary option (punish or not). By contrast, Smith and Warneken (2016) asked children to judge the amount of rewards or punishment hypothetical actors should receive for doing more or less of a good deed (e.g. cleaning windows) or bad deed (e.g. muddy footprints on the carpet). They found that 6- to 10-year-old children, but not 4- to 5-year-olds, assigned more aversive jobs to those who showed

more blameworthy behaviors. Thus, as they grow older, children allocate punishment in proportion to the amount of blameworthy behaviors, suggesting the development of desert-based punishment.

In sum, existing research shows that in situations involving helping or hindering one's instrumental goal, 2-year-olds already direct reward and punishment towards agents based on desert (Hamlin et al., 2011). By 3 years of age, children intervene against moral transgressions by reproaching verbally (Vaish et al., 2011) or by preventing access to activities after a property damage (Yudkin et al., 2020). Around 6 years, children systematically intervene against fairness norm violations, punishing unfair allocations more often than fair allocations (McAuliffe et al., 2015) and endorsing relatively more punishment to more severe norm violations (Smith & Warneken, 2016). Therefore, as children grow older, they not only intervene against an increasing range of transgressions — from goal hindrance, ownership violations to fairness violations — but also intervene in a systematic and selective manner.

Evaluations of Third-Party Punishers in Children

While the studies illustrated above show when children start to engage in punishment at a behavioral level, their understanding of punishment could emerge earlier than behavioral enactment of punishment. Previous research assessed infants' understanding and evaluations of a third-party punisher. For example, in an interaction involving physical aggression (e.g., a perpetrator chasing and hitting the victim), 6-month-old infants were more likely to touch a third-party agent who intervened by blocking the perpetrator from the victim over another third-party who did not intervene, demonstrating a preference for intervenors (Kanakogi et al., 2017). Another study (Hamlin et al., 2011) tested infants' preferences in a context that involves helping or hindering one's goal. When infants had to choose between a taker-puppet who had removed a

treat from a goal-hinderer and a giver who had handed a treat to the goal-hinderer, 8-month-olds, but not 5-month-olds, preferred the taker over the giver, suggesting that they like those who punish hinderers better than those who help hinderers.

In Vaish, Herrmann, Markmann and Tomasello (2016), 4- and 5-year-olds watched scenarios in which a transgressor broke a moral norm (e.g., one person destroying another person's belongings). Subsequently, one third-party enforced the norm verbally (e.g., "Don't ever do that again"), whereas the other third-party made neutral comments. Results showed that 5-year-olds, but not 4-year-olds, evaluated norm-enforcers more positively than non-enforcers. Hence, in a context of ownership violations, children's own spontaneous verbal protest emerges around 3 years (Vaish et al., 2011), whereas it is not until 5 years of age that children reflect on the behaviors of a third party and like those who verbally enforce norms more than those who do not. Taken together, the studies with infants and children suggest their preference for punishers over non-punishers in various forms of moral transgression.

These studies provide insight into children's evaluation of third-party punishers. However, what is not known is whether children choose punishers because they like them or merely because they are paired with more negatively valenced individuals. Specifically, punishers have been compared with bystanders who witnessed the transgression but chose not to intervene (e.g., Kanakogi et al., 2017, Vaish et al., 2016) or with givers who helped the transgressor (e.g., Hamlin et al., 2011). It is therefore not clear whether they evaluate punishers positively or just not as negatively as the alternative agent.

Moreover, prior research focused exclusively on moral transgressions such as hindering one's goal (e.g., Hamlin et al., 2011), physical aggression (e.g., Kanakogi et al., 2017) and property damage (e.g., Vaish et al., 2011; Vaish et al., 2016). Equally important to consider is

how children evaluate the punishment of transgressors when *fairness norms* are violated. As illustrated earlier, third-party punishment is an important mechanism for dealing with individuals who violate fairness norms. In this dissertation, I tested children from 5 to 9 years of age because this is the age range when children show significant development in their understanding of fairness norms. For instance, around 6 years of age, children enact punishment against those who violated fairness norms (McAuliffe et al., 2015). Between 7 and 8 years, children reject offers advantageous to themselves (Blake & McAuliffe, 2011), suggesting that children start to apply fairness norms in an impartial way.

Research Questions and Proposed Research

In this dissertation, I aim to answer three key research questions. First, when and how does third-party punishment emerges in childhood? Second, what motivates children to punish unfairness and what are the conditions under which punishment increases or decreases? Third, how does children's reasoning about punishment interact with an alternative form of intervention such as third-party helping? Ultimately, the focus on TPP would allow us to consider questions such as: what are the antecedents and consequences of third-party punishment in children?

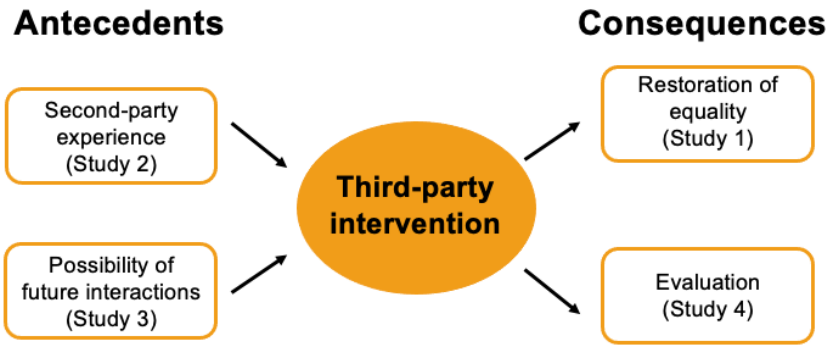

Figure 1. Schematic representation of the proposed research.

I propose four sets of studies to address these questions (see Figure 1). Study 1 investigates the distributional aim of third-party punishment in children. Specifically, this study tested whether children use third-party punishment to restore equality between two other individuals.

Study 2 encompasses two studies that examine whether children's experience as a second-party recipient (i.e., whether they received fair or unfair allocations) affects their subsequent third-party punishment.

Study 3 assesses whether the possibility of encountering the same divider in the future interactions would affect children's subsequent third-party punishment. Specifically, I tested whether children are more likely to punish unfair dividers when they are told that the same divider (vs. a new divider) will share resources with child themselves.

Study 4 investigates children's evaluations of third-party punishers. By comparing third-party punishers with third-party helpers, I aim to determine the extent to which third-party punishers receive positive reputations.

Chapter 2: Study 1

Third-party Punishment in Children Aims at Equality

Abstract

Third-party punishment has been regarded as an important mechanism to promote fairness. While previous research has shown that children aged 6 and older punish unfair behaviors at a personal cost, it is unknown whether they actually intend to establish equality or whether equality is a mere byproduct of punishment. In this pre-registered study, $N = 60$ 5-to-9-year-olds witnessed how an agent made unfair resource allocations to a peer. Children could then pay a personal cost to intervene and decide not only whether to punish, but also how much to punish. I found that with age, children calibrate the degree of punishment to equalize outcomes between third parties.

Third-party punishment in children aims at equality

Children's TPP against unfairness has been found in children aged 6 and older (Gummerum & Chu, 2014 with samples from the UK; House et al., 2020 with samples in Argentina, India, Germany and the US; Jordan, McAuliffe, & Warneken, 2014; McAuliffe, Jordan, & Warneken, 2015 with samples in the US). For example, in McAuliffe et al. (2015), children were shown how an absent child (hereafter divider) allocated 6 Skittles between the self and another absent child (hereafter recipient). The divider made either fair (3 for the self, 3 for the recipient) or unfair allocations (6 for the self, 0 for the recipient). As a third-party observer, children could accept the allocation, which meant that the Skittles were distributed the way the divider had allocated them, at no cost to the participant. The alternative was for the participant to pay one of their own Skittles to reject the divider's allocation. Then, all 6 Skittles were thrown away and became inaccessible to everyone. Therefore, rejection serves as a third-party punishment, resulting in a 0:0 outcome for the divider and the recipient. McAuliffe et al. (2015) found that 6-year-olds rejected unfair allocations more often than fair allocations. Whereas, 5-year-olds showed a similar, but less reliable pattern of punishment. This study suggests that TPP against unfairness develops around age 6.

Although these earlier studies provide insight into the development of TPP, one critical question remains. In previous studies, children's punishment was binary (e.g., House et al., 2020; Jordan et al., 2014; McAuliffe et al., 2015). For example, children could stay with the divider's original allocation by accepting it or remove everything (0:0) by rejecting it. Consequently, punishment automatically resulted in equality between two individuals (0:0). Therefore, it is unclear whether children genuinely aimed for equality when they chose to punish or equality was a byproduct of their punishment decisions. For example, it is possible that children punish selfish

dividers to see the person suffer or to avenge the victim without an intention to restore equality. If this was the case, the equality that resulted from punishment was not the main intention, but only a side-effect of the child's goal to punish.

Here I examine whether children punish with the aim to create equality. I start with the notion that at least in adults, TPP has been identified as a potential mechanism to enforce norms, including fairness norms. Therefore, if children's TPP is motivated by a norm of equality, children should punish in a way that reduces inequality among third parties. Alternatively, if their TPP is driven by a self-centered motive such as spite, competition (Fehr, Hoff, & Kshetramade, 2008; Raihani & Bshary, 2019) or a desire to watch the deserved punishment enacted (Mendes et al., 2018), the focus would be on inflicting costs on others, without consideration on whether punishment reduces inequality. To date, whether children use TPP to endorse the fairness norm has not been measured directly. The current study is a test of children's norm-based punishment by assessing the distributional end state their punishment pursues.

To examine whether children punish with the goal of creating equality, I presented participants with three (pre-programmed) allocations between two peers that were represented as two avatars on a computer screen: fair allocations (2:2), mildly unfair allocations (3:1) and extremely unfair allocations (4:0). Critically, our participants were in the role of a third-party and were free to choose how many coins they wanted to take away from which individual and therefore decided not only whether to punish, but also on the degree of punishment.

I tested the hypothesis that children use punishment to establish equality against several other possible outcomes. Specifically, based on findings that by school-age, children from the US gravitate towards equal sharing of rewards (Blake & McAuliffe, 2011; Shaw & Olson, 2012),

I predicted that between 7 and 8 years of age, children would become more likely to fine-tune their punishment to balance the scales between two third parties (e.g., turning 3:1 into 1:1). However, several alternative outcomes are plausible. For example, children might punish unfair allocations more often than fair allocations, but not yet be able to calibrate their punishment to restore equality (e.g., making 3:1 into 2:1). Another possibility is that children might be motivated to avenge the recipient by over-punishing the selfish divider (e.g., making 3:1 into 0:1) without considering how their punishment tilts the scales in the opposite direction. Another possibility is that children punish to deprive others of rewards with the competitive goal to end up with more resources than others. If this is the case, children should punish fair as well as unfair allocations, and should take all coins away from the recipient as well as the divider (e.g., making 3:1 into 0:0). Our study was designed to assess these different possibilities.

Method

Participants. Our final sample were $N = 60$ 5- to 9-year-old children ($M = 88.47$ months, range = 61 - 119 months, $n = 12$ in each age group, 30 male, 30 female). Children were tested at a museum in the Midwest of the US. Demographic information such as race, education and income could not be obtained as per the rules of the museum. Four additional children were excluded because of failure to correctly answer at least one of the comprehension checks (2), parental interference (1), or parental report of their child having autism (1). Power analyses established that our sample size was large enough to detect effects of interest.

Experimental design and procedure. After parents gave written consent, children sat at a table with the study apparatus while the parents watched passively from a few steps away. A female experimenter introduced the computer game referred to as the "coin game" and explained that players could collect virtual coins to later exchange for prizes. During a *prize introduction*,

children learned that the more coins they have during the coin game, the more and the better prizes they would be able to choose afterwards.

In the subsequent *practice phase*, the experimenter introduced the two other players in the game by stating that they were children of the same age and gender at another museum, who are currently connected online. In reality, the decisions of the two other players were computer-programmed. The experimenter introduced the role of the *divider* and the *recipient*: The divider could decide how to divide 4 coins between the self and the recipient. The divider could make one of three allocations: (a) 2 for the self and 2 for the recipient, (b) 3 for the self and 1 for the recipient and (c) 4 for the self and 0 for the recipient. The recipient was a passive player who could only accept the divider's allocation.

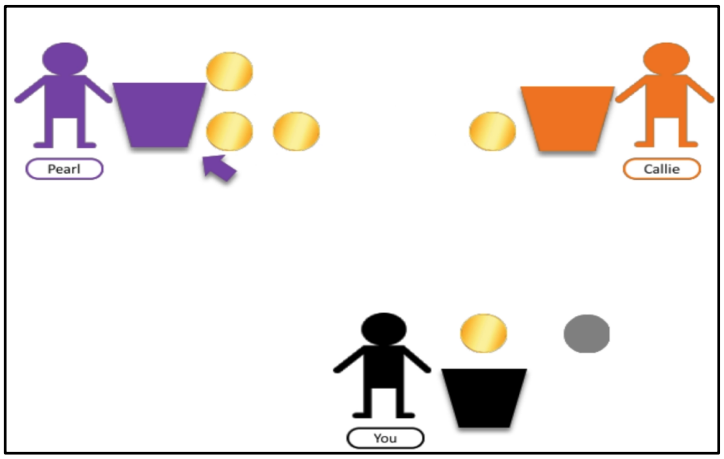

Figure 2. Computer screen display of the coin game during the test phase in Study 1.

This example shows a 3:1 allocation made by the divider (Pearl on the top left corner) to the recipient (Callie on the top right corner). Children as a third-party observer at the bottom could push either their own coin above the black basket (rejection) or the gray button (acceptance).

After introducing the roles, children watched on the screen how the divider produced different allocations and practiced their role as a third-party punisher. After the divider made an allocation, children could press either the gray button or their own coin above their basket

(Figure 2). If they pushed the gray button (acceptance), the four coins went into each player's basket just the way the divider allocated the coins, and the child's own coin went back into their own basket. That is, acceptances incurred no cost to the child.

Alternatively, if children pushed their own coin above their own basket (rejection), they could decide which other coins they wanted to take away from either the divider or the recipient or both players. When children pushed a coin, a vacuum appeared at the top of the screen and sucked up the coin, such that no one could keep the coin. Any remaining coins on the screen that children choose not to take away from players went into each player's basket. Therefore, in our game, children could be flexible about who to punish (divider, recipient or both) and the number of coins they want to take away (up to 4 coins), including taking all coins away from the other players. Critically, for children to punish others, they first had to sacrifice their own coin, making punishment costly for them.

There were four practice trials in total. Children practiced four possible outcomes of each button (accept vs. reject) in each allocation (fair vs. unfair). The experimenter asked comprehension checks about the consequence of each button and whether each button required the payment of the child participant's coin or not. If the child answered incorrectly, the experimenter corrected the answer and asked the question(s) again in the next practice trial. All 60 children included in the data analysis passed these comprehension checks.

After the four practice trials, to make children believe that the other players were real, the experimenter pretended to call the other players on speakerphone and checked if they were ready to play the game. In reality, a confederate answered the phone call. Upon the completion of the study, the experimenter left and a secondary experimenter asked children whether they thought

the players were real or pretend. I found that 78% of children (47 out of 60) said the players were real.

During the subsequent *test phase*, children played 6 rounds as a third-party observer in total. Children received 20 coins as their initial endowment (coins that dropped into their basket on the screen). The divider and recipient were different from those in the practice trials (with different names). Each child played 2 rounds of 2:2 allocations, 2 rounds of 3:1 allocations and 2 rounds of 4:0 allocations, presented in a pseudo-random order with the restriction that two identical allocations were not presented consecutively. In each round, children could press one of two buttons (gray or their own coin) to either accept or reject the divider's allocation towards the recipient. The role of each player remained the same throughout the test trials.

After the test phase, to assess whether children's numeric understanding affected their performances in the coin game, I administered a non-social numeric task. Children saw four blue triangles on the computer screen that were identical to the three allocations in the coin game and were asked to make both sides equal. Children were able to create equal numbers of triangles in 96% of the trials. Thus, any potential age-related sensitivity to equality in the coin game cannot be attributed to their numeric inability to make both sides equal.

I counterbalanced the order of test trials, practice trials, deception check questions, numerical task trials, the other player's identity, and the type of unfair allocation during practice.

Data coding and analyses. Children's responses were automatically recorded by GameMaker Studio (https://www.yoyogames.com) and later checked and entered into a spreadsheet by independent coders. All statistical analyses were conducted with R statistical software (R version 3.5.2; R Core Team, 2018).

I pre-registered our hypotheses and analyses before data collection (https://aspredicted.org/blind.php?x=3iu8ci). All data and protocols are available through the Open Science Framework: https://osf.io/3v8zw/?view_only=778850243242493dae24099b8d9e58e7

I analyzed the difference in the number of coins between the divider and recipient with a linear regression, and rejection rate and the rate of creating equality with Generalized Linear Mixed Models (GLMM). I compared a full GLMM, which included age in months and allocation and an interaction between age and allocation as fixed effects and subject ID as a random effect with a null model, which included only subject ID as a random intercept. If the full model provided a significantly better fit to the data, I created a minimal model by sequentially dropping single terms from the full model, and finalized our minimal model when dropping single terms no longer provided a better fit to the data.

Results

When do children punish? I first assessed whether children's decision to punish was influenced by age and allocation. A full GLMM on children's rejection (0 = acceptance, 1 = rejection) provided a significantly better fit to the data than the null model (LRT, χ^2 (5) = 36.11, $p < .001$), further revealing a significant interaction between allocation and age (LRT, χ^2 (2) = 22.46, $p < .001$).

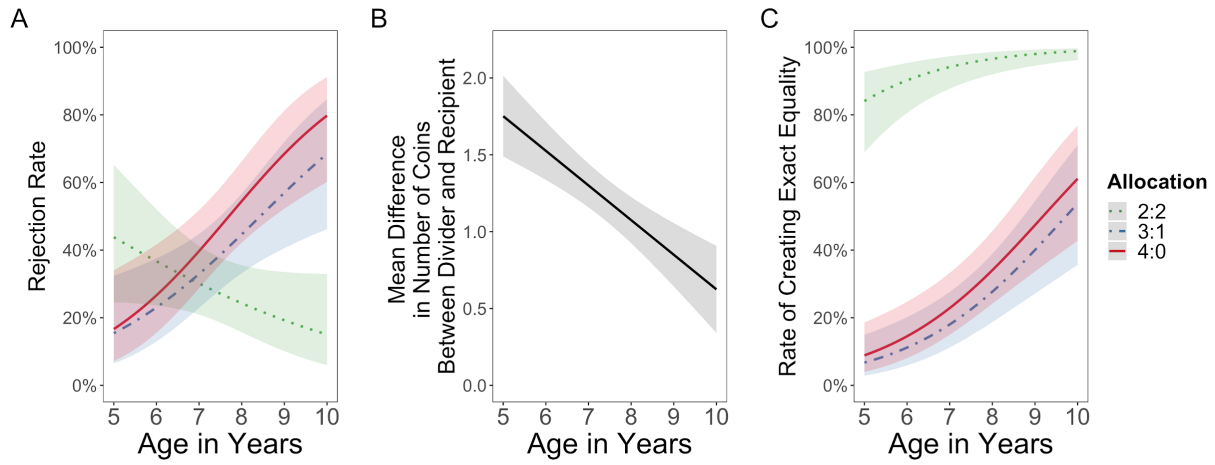

Figure 3. Children's punishment decisions in Study 1.

(A) Estimates of children's rejection rate with 95% confidence intervals based on the final model. (B) Estimates of the mean difference in one trial with 95% confidence intervals based on the linear regression. In the Y axis, 2 indicates that, on average, the divider had 2 more coins than the recipient in a trial after children's intervention. (C) Estimates of the rate of creating equality with 95% confidence intervals based on the final model.

As shown in Figure 3A, estimates for children's rejection rate of unfair allocations increased with age from around 20% of trials at age 5 to over 70% of trials by age 9 (LRT, χ^2 (1) = 9.20, b = 0.04, SE = 0.01, p < .01 for 3:1 and χ^2 (1) = 13.15, b = 0.05, SE = 0.02, p < .001 for 4:0), whereas rejection of fair allocations remained low at around 20% of trials overall (and even tended to slightly decrease with age; LRT, χ^2 (1) = 3.32, b = -0.03, SE = 0.02, p = .07). In fact, pairwise comparisons showed that the trajectories of responding to unfair allocations differ both between 3:1 and 2:2 (b = 0.07, SE = 0.02, p < .001) and 4:0 and 2:2 (b = 0.07, SE = 0.02, p < .001), but not from each other (4:0 vs. 3:1, b = 0.01, SE = 0.02, p > .62). In sum, with age, children became more likely to reject unfair over fair allocations and rejected allocations of 3:1 and 4:0 at similar rates. These results are consistent with the previous work (McAuliffe et al., 2015) that with age, children are more likely to punish unfair over fair allocations.

Do children punish to reduce inequality? To assess our main question about children's *degree* of punishment, I first examined whether children were more likely to reduce inequality through punishment with age. To assess this, I measured the difference in the number of coins between two players at the end of each trial and averaged it across all trials, resulting in one mean difference score per child, with higher values indicating greater inequality while a score of zero indicating perfect equality.

Estimates of the mean difference in the number of coins between two players declined from 1.75 coins at age 5 to 0.62 coins at age 10. Our analyses showed that with age (b = -0.02, SE = 0.003), the difference in the number of coins between the divider and the recipient decreased after children's intervention, R^2 = .26, $F(1, 58) = 22.08$, $p < .001$ (Figure 3B). Hence, as children grow older, they use punishment as a way to minimize inequality between individuals.

How often do children create exact equality? The result from the mean difference showed that the outcome of children's punishment gets closer to equality with age. Additionally, I examined the rate at which children used punishment to establish perfect equality such that both players ended up with the exact same number of coins. This analysis using success or failure at reaching exact equality as a binary variable is stricter because it tests whether punishment perfectly hits the target outcome (equality).

A full GLMM on children's creation of perfect equality between two other players (0 = inequality, 1 = equality) provided a better fit to the data than the null model (LRT, χ^2 (5) = 183.39, $p < .001$). I found significant main effects of allocation (LRT, χ^2 (2) = 163.51, $p < .001$) and age (LRT, χ^2 (1) = 19.31, $p < .001$), but no interaction between age and allocation (LRT, χ^2 (2) = 2.92, $p > .23$). Pairwise comparisons from the main effect of allocation revealed that exact

equality occurred more often after fair allocations (M = 0.93, SD = 0.26) than in unfair allocations (b = -4.29, SE = 0.52, $p < .001$ for 2:2 vs. 3:1; b = -3.99, SE = 0.51, $p < .001$ for 2:2 vs. 4:0), which is not surprising given that equality already existed in fair allocations.

More importantly, I found a main effect of age, suggesting that across allocations, children were more likely to establish exact equality with age (b = 0.05, SE = 0.01, $p < .001$). As can be seen in Figure 3C, estimated rates of establishing perfect equality in unfair allocations increased from below 10% of trials at age 5 to over 50% of the trials at age 10. These rates of creating equality did not differ from each other between 3:1 (M = 0.26, SD = 0.44) and 4:0 allocations (M = 0.31, SD = 0.46; b = 0.30, SE = 0.32, $p > .34$). This analysis provides converging evidence that with age, children become more likely to create equality between third parties.

How many coins do children take away in each allocation? Our final question was how many coins children took away from which individual when they decided to punish. To assess this, I calculated the average number of coins taken away from each player in the 141 trials (39% of test trials) in which children chose to reject the allocation (Figure 4). In terms of the number of coins children took away from the divider, they took 3.54 out of 4 coins in 4:0 allocations, 2.25 out of 3 coins in 3:1 allocations, and 1.54 out of 2 coins in 2:2 allocations. In terms of the number of coins children took away from the recipient, in 3:1 allocations, 5-year-olds took 0.80 out of 1 coin. However, children tended to take less coins away from the recipient with age: 9-year-olds took 0.07 coins away from the recipient. In 2:2 allocations, children took 1.68 out of 2 coins away.

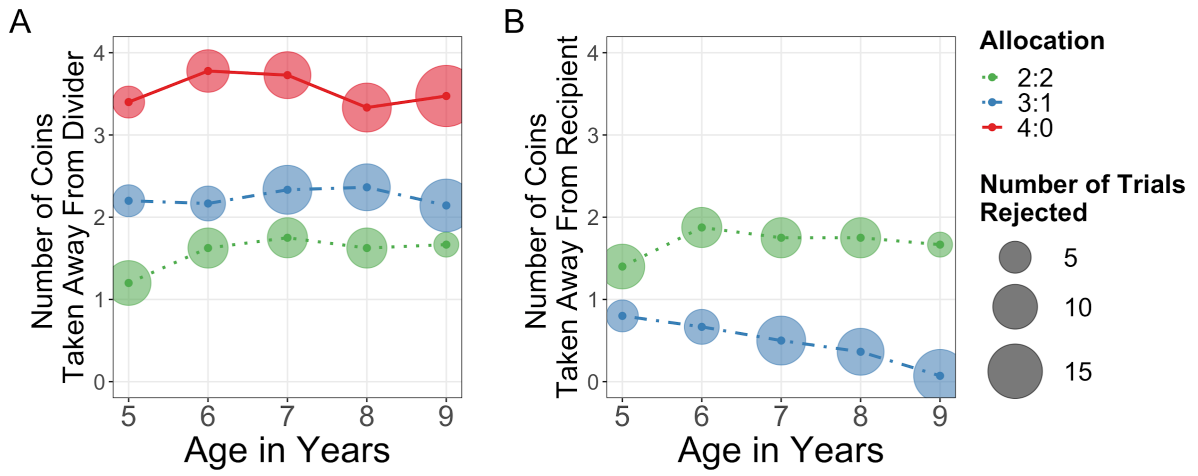

Figure 4. Number of coins taken away from players in Study 1.

(A) Average number of coins taken away from the divider in each age group. (B) Average number of coins taken away from the recipient in each age group. The size of circles represents the frequency of rejection. In Figure 4B, 4:0 allocations are not included because in these trials the recipient had no coins that could be taken away.

Overall, when children rejected unfair allocations, they took coins from the divider in a way that established equality and were less likely to take coins from the recipient with age. When children rejected fair allocations in a minority of trials, they took coins away from the recipient as well as the divider.

Discussion of Study 1

Our findings demonstrate the development of fairness norm-based punishment in which children use punishment to create equality. This was possible because in our task, children were able to express their distributional preference by deciding not only whether to punish, but how much to punish. I showed that with age, (a) children were willing to pay a cost so that punishment would reduce inequality between individuals, and (b) they became increasingly

better at creating exact equality. These findings support our hypothesis that with age, children fine-tune their punishment to restore equality.

Our findings rule out several alternative hypotheses of what could have motivated children's punishment. If children's punishment had been primarily driven by self-centered motives such as competition or inflicting costs on others, they would not have cared about creating equal outcomes between two other individuals. For instance, if they had been motivated to win against others, children should have taken all coins from not only the divider but the recipient in 3:1 allocations. However, this was not the case: 9-year-olds took about 2 coins from the divider and almost 0 coins from the recipient, suggesting that their punishment targeted specifically at the selfish divider. Similarly, if children had been concerned exclusively about avenging the victim, their punishment would have ended up in another unequal outcome that favors the victim over the divider (e.g., making 3:1 into 0:1), which was not the case. Overall, our findings show that children's costly punishment is motivated by a concern for fairness, with children becoming increasingly more likely to use punishment to create equality between two individuals.

Our results show important age effects. Children at age 5 punished fair allocations at a similar rate as unfair allocations. However, with age, children's tendency to punish fair allocations declined, while their willingness to punish unfair allocations increased. Nonetheless, even though 9-year-olds as the oldest tested age group were most likely to create equality, they were far from perfect (establishing perfect equality in 46% of 3:1 allocations and 54% of 4:0 allocations). One major reason is that they did not always punish unfair allocations, perhaps because punishment was costly for them and they were hesitant to always punish even though they were capable of establishing equality.

These results add to the existing literature that children around age 8 and older give up their own resources to avoid getting more than others, which suggested that children are averse to inequality even when it is advantageous to themselves (Blake & McAuliffe, 2011; Shaw & Olson, 2012). Furthermore, our findings are consistent with a recent study that children's punishment targets unequal outcomes regardless of the divider's intention behind the outcome (Bernhard, Martin, & Warneken, in press). Together, children's focus on unequal outcomes implies that the egalitarian motive to reduce differences in payoffs among people could underlie children's punishment, as had been previously shown with adults (Dawes et al., 2007; Fehr & Schmidt, 1999; Johnson et al., 2009). Importantly, our study shows that children calibrate the amount of punishment not only to prevent unequal outcomes, but specifically to create allocations that move others closer to equality.

Although children showed an increased sensitivity to equality with age, there were noteworthy patterns observed in young children. For example, 5-year-olds were more likely than older children to take coins away from the disadvantaged recipient (Figure 4). However, the lack of equality restoration in young children cannot be attributed to their lack of numerical understanding because children were able to make both sides equal in our non-social numeric task. Also, our findings do not necessarily imply that 5-year-olds lack a sense of fairness per se. For example, in contexts in which there was no cost to the self, infants and young children demonstrate a preference for equal over unequal allocations (Cooley & Killen, 2015; Fehr et al., 2008; Geraci & Surian, 2011; Sloane, Baillargeon, & Premack, 2012). Instead, I speculate that 5-year-olds were in the process of learning when and how much to punish in a context that involves a cost to the self.

One potential concern is that the use of a computer game limited children's comprehension of the social interaction. That is, the advantage of allowing for better experimental control and anonymity in a computer game with actual costs to the child might be offset by having to simulate a social interaction. To address this concern, the task was designed to be highly engaging for children and included comprehension checks showing that children had no difficulty following the game. However, while the vast majority affirmed our manipulation check that they interacted with real children, 13 out of 60 did not. To address this, I confirmed that the patterns of results were the same when these children were excluded from analyses, suggesting that they did not bias our findings. Another concern is that because children interacted with the same divider and recipient throughout, they might not have focused on individual allocations but the whole sequence of events. However, supplementary analyses indicated that there were no effects involving trial number or children's rejection in a previous trial.

One question for future research is *why* older children showed equality-oriented TPP. It is possible that this reflects a genuine sense of fairness that emerges at this age. However, it is also possible that this is in part influenced by older children wanting to signal how much they care for fairness. It is known that children from age 8 show a clear sign of reputation management (Engelmann & Rapp, 2018). Therefore, reputational concerns might partly influence equality-oriented TPP in children aged 8 and older. Future research should examine potential motivations underlying TPP in children.

Chapter 3: Study 2

The Influence of Age and Experience of (Un)Fairness on Third-party Punishment in Children

Abstract

Third-party punishment is an important mechanism to enforce norm-following. Previous research has shown that by around 6 years of age, children begin to punish individuals who violate fairness norms, even when they are an unaffected third-party and have to pay a cost to intervene. However, the underlying process that explains the development of third-party punishment is poorly understood. Here I examine to what extent age-effects and contemporaneous experiences of receiving unfair offers influence third-party punishment. In two studies, a total of $N = 280$ 5- to 9-year-olds played a computer game in which they received either fair or unfair offers of coins from another player. In the subsequent test phase, children could intervene against unfair offers as an unaffected third-party. Across both experiments, I found that with age, children become increasingly more systematic in their decisions to intervene against unfair, but not against fair allocations. However, there was no strong evidence that an immediate experience of (un)fairness influenced children's subsequent decisions. Together, our results suggest that children develop a sophisticated application of fairness norms with age that is not easily swayed by concurrent experiences. I discuss how these results contribute to the literature on the development of fairness.

The influence of age and experience of (un)fairness on third-party punishment in children

TPP against distributional unfairness develops later in children compared to other types of moral transgressions. Children show the first sign of TPP around 5 years of age and demonstrate a reliable punishment from 6 years (Gummerum & Chu, 2014; House et al., 2020; Jordan, McAuliffe, & Warneken, 2014; McAuliffe, Jordan, & Warneken, 2015). In one study (McAuliffe et al., 2015), 6-year-olds show reliable TPP, punishing more often in unequal allocations than in equal allocations. Hence, even though they were an uninvolved third-party, 6-year-olds reliably inflicted a cost on the selfish divider at their own cost whereas, 5-year-olds showed less reliable rates of punishment. This study suggests that costly TPP against unfairness develops between 5 and 6 years in children from the US (see House et al., 2020 for cross-cultural variations in children's TPP).

These results raise the question about the underlying process that explains this developmental trajectory. One potential reason is that older children are better able to take the perspective of the disadvantaged child. Under this hypothesis, older children are more adept at imagining what the victim of unfairness is experiencing and feel compelled to set things straight after they put themselves in the victim's shoes. This would be consistent with the notion that we rely on *simulation* to predict others' minds and behaviors (Gordon, 1986; Harris, 1992). That is, one's own mind can be used as a basis or a model for understanding other people (e.g., "What would I do if I were in the same situation?"). Importantly, the notion here is that an understanding of another person's psychological world starts with one's own experience. Therefore, having a similar experience in the past should be able to enrich the simulation process, leading to a better understanding of others in the same circumstance. Applied to the fairness context, children's personal experience of unfairness could increase their sensitivity to a

third-party's unfair sharing by allowing them to apply their experience and to simulate others' minds (e.g., "What would I want to do if I were treated unfairly?").

Furthermore, prior work has hinted at the importance of children's direct experience in their sense of fairness. These studies have shown how children react when they receive unfair sharing. For example, when children receive a smaller amount of resources than a peer partner, children as young as 4-years-old protest against the unfair allocation by choosing to receive none instead (Blake & McAuliffe, 2011; Blake et al., 2015). Moreover, by directly comparing second- and third-party contexts, Bernhard, Martin, and Warneken (in press) found that 5- and 7-year-olds are more likely to punish someone who treated them unfairly in a second-party context than when they witnessed how someone else was treated unfairly in a third-party context. These studies, therefore, suggest that the personal experience of receiving unfair treatment is perhaps primary in children's desire to intervene and the developmental shift towards TPP consists of becoming able to apply this first-hand experience to others. One way to test this would be to provide children with the direct experience of unfair treatment as a facilitator of TPP. Building upon this idea, I hypothesized that children who received unfair allocations from others would be more likely to intervene against unfairness as a third-party compared to those who did not receive unfair allocations.

An alternative hypothesis to this simulation account would be that as they grow older, children develop a more principle-based and perhaps impartial fairness concept. The idea here is that children do not necessarily rely on simulation by extending their own experiences. Instead, with age, they acquire and apply general fairness principles in a more unbiased, agent-neutral way, in which an interest or perspective of an agent (including children themselves) is not favored over that of another agent.

Several pieces of evidence show that this hypothesis is plausible. For example, around age 8, children from the US give up their own resources to avoid getting more than others, showing an aversion to inequality advantageous to themselves (Blake & McAuliffe, 2011; Shaw & Olson, 2012). Further, 8-year-olds, but not 6-year-olds, punish unfair dividers not only when their ingroup members are harmed but also when outgroup members are harmed (Elenbaas, Rizzo, Cooley, & Killen, 2016; Jordan et al., 2014). Together, these studies suggest that with age, children apply fairness norms to the self and ingroup members as well as others and outgroup members, impartially adhering to the norm. This shows that children develop general principles that dictate their sense of distributional justice.

Under this approach, the prediction would be that children's own experience would not have a major influence on their subsequent TPP. Instead, regardless of their immediate personal experience, as children grow older, they develop general principles that guide how resources should be distributed and what is the right thing to do in the face of unfairness.

Current Study

The current study tested the impact of children's previous sharing experiences on their subsequent willingness to punish unfairness at a personal cost. To manipulate children's sharing experiences with other people, I developed a novel computer game, in which child participants have live interactions with other (computer-programmed) players and watch how they divide coins that can be exchanged for prizes afterward. Children were assigned to one of three conditions: (a) unfair experience condition, in which they always received unfair offers (0 out of 6 coins) from another player, (b) fair experience condition, in which they always received fair offers (3 out of 3 coins) from another player, and (c) no experience condition, in which they did not play the role as a recipient at all. In the subsequent test phase, children played a third-party

punishment game, in which children as a third-party could observe how one player divides coins with another player and decide whether to punish the offer or not.

I hypothesized that children in the unfair experience condition will be more likely to punish unfair offers as a third-party compared to those in the no experience or fair experience conditions. This hypothesis is built on the idea that experiencing unfairness at first hand would allow children to better simulate others' perspectives and thus heighten their sensitivity to a third-party's suffering.

One alternative hypothesis is that children will show a similar rate of TPP across conditions, suggesting no influence of second-party experiences. Instead, with age, children become more reliable in punishing unfair over fair offers, regardless of their own prior experience. If I find results consistent with this prediction, it would imply that children develop general principles regarding fairness with age that motivates their intervention, and at least immediate personal experiences do not alter their motivation.

Another potential outcome is that children will be *less* likely to intervene against unfairness after they have experienced unfairness themselves. While this might seem unlikely at first sight, there are two reasons why such an outcome should be considered. First, through their repeated first-hand experience of being treated unfairly, children might learn that people are selfish in the game and accept this as typical or normative behavior. Second, it is possible that children, who lost out by never being shared with are focused on the small amount of resources they ended up with, are therefore more hesitant to pay a cost to punish in a third-party context.

General Method

Experimental design and procedure. In both experiments, a female experimenter introduced the computer game referred to as the "coin game" to a child participant. I first

established that children could collect virtual coins that they could later exchange for prizes. Specifically, during a *prize introduction*, children saw an image of the prizes and the number of coins needed to purchase them later. The prizes and coins were shown as three tiers, illustrating that the more coins children have during the coin game, the more and the better prizes they could choose afterwards. For example, with a few coins, children could only have a sticky paper (lowest tier), but if they had a large number of coins, they could have a slinky, a ring, and a sticky paper (top tier). The prize introduction concluded with a comprehension check question about the prize hierarchy where the experimenter asked children to identify prize(s) they can have with the amount of coins at the lowest or the top tier. Children were correct in these comprehension check questions (98%) confirming their understanding of the exchange value of coins in the computer game.

In the subsequent *practice phase*, the experimenter introduced the two other players in the game by saying that they were children of the same age and gender as the participants who are currently connected online but are at another location. For example, children were told that the other players are at another museum (for those who were tested at a museum) or at another park (for those who were tested at a park). However, in reality, the other players were computer-programmed. The experimenter then introduced the role of the *divider* and the *recipient*. The divider can decide how to divide 6 coins between the self and the recipient. The divider can make one of two offers: equal offer (3 for the self, 3 for the recipient) or unequal offer (6 for the self, 0 for the recipient). The recipient was a passive player who can only accept the offer made by the divider.

After introducing the roles, children saw the offers made by the divider enacted on the screen and practiced their role as a third-party punisher. The experimenter told children that after

the divider makes an offer to the recipient, they can press either the green button or the red button. If they push the green button (acceptance), the 6 coins will go into each player's basket just the way the divider allocated the coins (e.g., if the divider splits it up 3:3, each player's basket receives 3 coins), and the child's own coin goes back into their own basket. That is, acceptances incurred no cost to the child.

In contrast, if children push the red button (rejection), a vacuum will appear at the top of the screen and the 6 coins will be sucked up and disappear into the vacuum. Critically, to enact the rejection, children first had to pay their one coin into the vacuum. Therefore, pressing the red button serves as a costly third-party punishment.

There were four practice trials in total. Children practiced four possible outcomes of each button (accept vs. reject) in each offer type (equal vs. unequal). The experimenter asked comprehension check questions about the consequence of each button and whether each button requires the payment of the child participant's coin or not. When the child answered incorrectly, the experimenter corrected the answer and asked the question(s) again in the next practice trial. A majority of children (98%) answered correctly to the comprehension check questions. I excluded 6 out of 299 children (2%) who answered incorrectly in at least one of the comprehension check questions in the last practice trial from the data analysis.

After the four practice trials, to make children believe that the other players are real, the experimenter pretended to make a phone call to the other players on the speakerphone and check if they are ready to play the game. In reality, another experimenter (confederate) answered the phone call, and there were no other players.

Children received 20 coins as their initial endowment (coins that dropped into their basket on the screen).

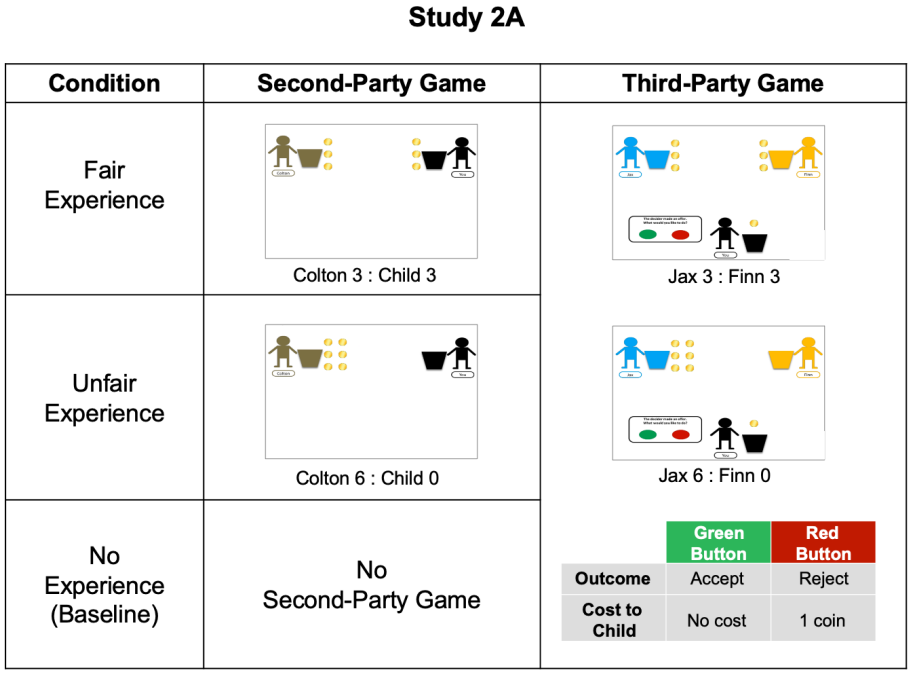

Figure 5. Schematic representation of the design in Study 2A.

In this example, Colton (left) was a divider and the child participant (right) was a recipient during the second-party game (experience phase). In the following third-party game (test phase), Jax (top left) was a divider and Finn (top right) was a recipient. The child participant (bottom) was a third-party observer. Regardless of conditions they were assigned, every child saw three equal offers (3:3) and three unequal offers (6:0) during the third-party game.

For the *experience phase*, children were assigned to one of three conditions (between-subject): *fair experience* condition, *unfair experience* condition, or *no experience* condition (see Figure 5). Children played the role of a passive recipient for four consecutive rounds during the second-party game with another (virtual) child as the divider. In the fair experience condition, the divider always kept 3 coins for the self and gave 3 coins to the child participant. In contrast, in the unfair experience condition, the divider always kept 6 coins for the self and gave 0 coins to the child. Those in the no experience condition (baseline condition) were unaware of the existence of the second-party game and did not play this game at all. After the first and the fourth round, the experimenter asked children to recall the number of coins they received from the

divider during the second-party game. Most children (99%) reported the number of coins they received correctly.

During the subsequent *test phase*, children played 6 rounds of the third-party game where the child was in the role of the third-party and two other children were the divider and the recipient, respectively. Critically, however, the divider and recipient in the third-party game differed from those in the practice trials or those in the second-party game to prevent potential carryover or retaliation towards the same divider. Within the test trials, the role of each player remained the same (e.g., Jax was always the divider, while Finn was always the recipient). 96% of the time, children correctly identified players from the second-party game and those from the third-party game. This shows that children were not confused and were aware of who had played each game.

There were 3 rounds with equal offers (3:3) and 3 rounds with unequal offers (6:0) presented in a pseudo-random order with a restriction that no more than two identical allocation types are presented consecutively. In each round, children could press one of two buttons (green or red) to either accept or reject the divider's allocation towards the recipient (see Figure 5). Importantly, children were unaware of the total number of rounds in the third-party game to prevent them from calculating the number of coins they currently have and the number of remaining rounds, which could influence their decision whether to spend a coin or not in a given round. Our dependent measure was children's rate of rejection, with rejection (red button) coded as 1 and acceptance (green button) as 0.

After the test phase, the experimenter left and a secondary experimenter asked children whether they think the players in the game are real or pretend players. Across two experiments,

91% of children (106 out of 120 in Study 2A, 149 out of 160 in Study 2B) reported that the players were real.

In both experiments, I counterbalanced the trial order of offer types during the third-party game, player's identity (names), left/right position of buttons (green button on the left vs. red button on the left), the order of practice trials (practice equal offer first vs. practice unequal offer first), and the order of comprehension check questions. Additionally, in Study 2B, I counterbalanced the color of boxes used in the winner and loser conditions.

Data coding and statistical analyses. Children's responses were automatically recorded by the computer game program, GameMaker Studio (https://www.yoyogames.com/gamemaker), and later checked and entered into a spreadsheet by independent coders. All statistical analyses were conducted with R statistical software (R version 3.5.2; R Core Team, 2018).

In both experiments, I analyzed children's rejection rate with Generalized Linear Mixed Models (GLMM) using the package *glmmTMB* (Brooks et al., 2017). In Study 2B in which I assessed children's expectations about offers, I analyzed the change in children's expectations about the divider's offer with linear models.

Our analysis procedure was as follows: (1) I examined a null model which included only subject ID in mixed models; (2) I created a full model which included our main predictors (e.g., age in months, condition, offer type) and all interactions among the predictors; (3) I compared the full model with the null model; (4) if the full model provided a significantly better fit to the data, I created a minimal model by sequentially dropping single terms from the full model, testing whether their inclusion improved the model fit; (5) I stopped this process and finalized our minimal model when dropping single terms no longer provided a better fit to the data.

As a supplementary analysis, I employed Bayesian statistics to provide more information about the robustness of our findings. A Bayes factor (BF) quantifies the degree to which the data favors the null hypothesis over the alternative hypothesis, and vice versa (Aczel et al., 2018; Wagenmakers, Morey, & Lee, 2016). Conventionally, a BF between 1 and 3 indicates anecdotal evidence, a BF greater than 3 suggests moderate evidence, and a BF greater than 10 provides strong evidence in favor of one hypothesis over the other (Jeffreys, 1961).

I computed BFs by comparing a GLMM in which a predictor of interest was included with a GLMM in which the predictor was not included, using the package *brms* (Bürkner, 2017). As in main analyses, I included subject ID as a random intercept in these models. The population-level regression coefficients had a weakly informative Student's t distribution prior which was zero-centered with 3 degrees of freedom and a scale of 2.5 (Gelman, Jakulin, Pittau, & Su, 2008). All models were run with 10,000 iterations with the first half as burn-in. $\hat{R}$ was less than 1.01 for all parameters, suggesting convergence (Vehtari et al., 2019).

I pre-registered both experiments before data collection. Pre-registered documents are available from AsPredicted: https://aspredicted.org/blind.php?x=fy77e4 for Study 2A and https://aspredicted.org/blind.php?x=j33fi6 for Study 2B. All data and protocols are available through the Open Science Framework: https://osf.io/upexj/?view_only=e3d5dea53d3c48e9b071580d82c2668a.

Pilot Study

In our pilot study, I tested $N = 32$ 5- and 6-year-old children in a laboratory with fair experience and unfair experience conditions. I chose this age because previous research had suggested that children around 5 to 6 years show costly TPP (McAuliffe et al., 2015). I piloted our aforementioned procedure, where children first experienced allocations as a second-party

recipient and then played the third-party game as a third-party punisher. This pilot study established that children were able to understand the computer game.

I found that children punished unequal offers more often than equal offers (LRT, χ^2 (1) = 28.48, b = 1.44, SE = 0.29, $p < .001$). Furthermore, children punished more often after an unfair experience than after a fair experience (LRT, χ^2 (1) = 4.28, b = 1.09, SE = 0.54, $p < .05$). After this validation, I made minor modifications to streamline the procedure and used this method in Study 2A and 2B. For these experiments, I broadened the age range to 5- to 9-year-olds to examine potential developmental change.

Study 2A

Method

Participants. Our final sample were N = 120 5- to 9-year-old children (M = 89.32 months, range = 60 - 119 months, n = 24 in each age group, 40 per condition, 60 male, and 60 female). Children were recruited and tested at museums or public parks in a city in the Midwest of the US. After parental consent, children were brought to a table with the study apparatus while the parents watched from a few steps away, instructed not to intervene. Demographic information such as race, education, and income was not assessed in this context.

Ten additional children were excluded because they failed to answer correctly in at least one of the comprehension check questions about the consequences and costs involved in each button until the last practice trial (4), there was interference from a parent, sibling or a friend (4), the child was not an English-speaker and was unable to understand the instruction (1) or the parent declined to provide the child's date of birth (1).

Experimental design and procedure. I used the same basic design and procedure described in the General Method with an additional measure. After the completion of the third-

party game, the experimenter asked children how close they feel to the recipient in the third-party game. The perceived closeness to the recipient was assessed based on the prediction that children in the unfair experience condition would feel closer to the recipient in the third-party game compared to children in the fair experience condition.

I adapted the Inclusion of the Other in the Self (IOS) task developed by Aron, Aron, and Smollan (1992). Upon the completion of the study, the experimenter showed children five pairs of overlapping two circles with varying degrees. In each pair of circles, one of the circles represents the child participant with the figure from the third-party game and the other circle represents the recipient from the third-party game with his or her figure. The perceived closeness was represented as the degree of overlap between the two circles (from no overlap for no perceived closeness to almost complete overlap for highest closeness).

Other than predicted, I found no differences in perceived closeness across three conditions. In this 5-point scale of perceived closeness, regardless of conditions, children tended to choose the pair with a medium degree of perceived closeness (No experience $M = 3.10$, $SD = 1.37$; Fair experience $M = 3.13$, $SD = 1.47$; Unfair experience $M = 3.10$, $SD = 1.32$). Therefore, I did not include perceived closeness as a predictor in our full model.

Results

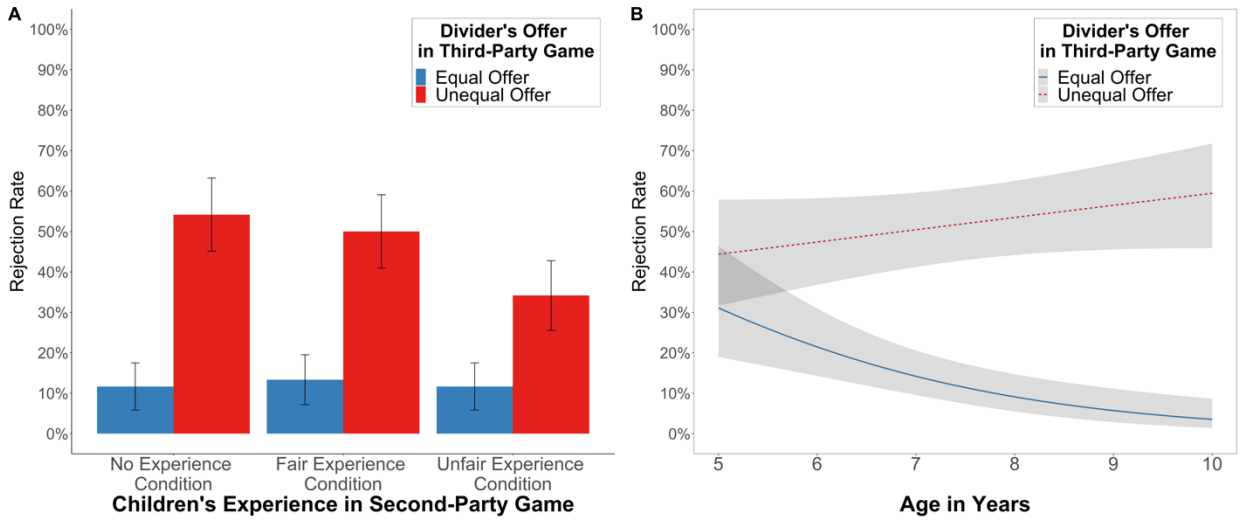

Figure 6. Children's punishment decisions in Study 2A.

(A) Children's rejection rate by condition in Study 2A. (B) Estimates of rejection rate based on the final model in Study 2A (collapsed across conditions). Error bars and confidence bands represent 95% confidence intervals.

Rejection rate. A full GLMM on children's punishment decision (0 = acceptance, 1 = rejection) with condition, offer type, age, and interactions among the predictors as fixed effects and subject ID as a random effect provided a significantly better fit to the data than the null model with only random intercepts (LRT, χ^2 (11) = 139.69, $p < .0001$).

Our critical question was whether children's TPP varies depending on their experiences of (un)fairness. Results revealed a significant main effect of condition, LRT, χ^2 (2) = 7.24, p = .03 (and no two- or three way interactions). Specifically, children in the unfair experience condition (M = 0.23, SD = 0.42) punished significantly *less often* than those in the no experience condition (M = 0.33, SD = 0.47; b = 0.64, SE = 0.26, p = .01) or fair experience condition (M = 0.32, SD = 0.47; b = 0.55, SE = 0.26, p = .03, see Figure 6A). By contrast, children in the no experience and fair experience condition did not differ from each other in their punishment rate

(b = 0.09, SE = 0.25, p > .72). These results suggest that the experience of unfairness decreased children's punishment, while the experience of fairness had no effect relative to baseline.

Another important question concerned whether children's TPP changes with age. I found a significant interaction effect of age and offer type (LRT, χ^2 (1) = 18.69, p < .001). To unpack this interaction, I ran separate models for equal and unequal offers. The results indicated that with increasing age, children become less likely to punish equal offers (LRT, χ^2 (1) = 14.51, b = -0.72, SE = 0.20, p < .001; see Figure 6B). However, there was no such age-related change in children's punishment of unequal offers (LRT, χ^2 (1) = 1.98, b = 0.19, SE = 0.13, p > .15). That is, their punishment of unequal offers remains relatively high across our age-groups but that of equal offers declines with age. Another question is at what age do children begin to reliably engage in costly TPP. Inspection of confidence intervals revealed that from 69 months of age (5.75 years), the confidence interval of equal offers no longer overlaps with that of unequal offers. Taken together, these findings suggest that (1) costly TPP develops between 5 and 6 years, and (2) as they grow older, children become more selective about the target of their punishment, understanding better when to punish and when not to punish.

To examine whether our results provide strong support for the effect of condition reported above, I computed Bayes factors (BF) for effects involving condition[1]. This revealed moderate evidence in favor of an *absence* of the main effect of condition (BF_{01} = 4.47), suggesting that the data is about 4 times more likely to be observed under the hypothesis that children's prior experience has no effect than the hypothesis that prior experience has an effect

[1] As Bayes factors are sensitive to prior distributions, I examined four other prior distributions. The results with different priors suggest that our BFs are relatively robust against various prior distributions. Even with the exploration of four other priors, I did not find evidence supporting the effect of experience of (un)fairness on TPP.

on punishment. Furthermore, because I predicted that children will intervene more often in unequal offers than in equal offers after experiencing unfairness, I computed a BF of an interaction between condition and offer type, which provided evidence in support of an absence of the interaction effect (BF_{01} = 8.93). In contrast, I found very strong evidence for an interaction effect between age and offer type (BF_{10} = 805), indicating that the data is 805 times more likely to be observed under the hypothesis predicting an interaction effect than under the hypothesis that there is no such interaction.

Taken together, results with BFs suggest that there is strong support for the finding that children intervene differently depending on their age. While children in the unfair experience condition were slightly less likely to punish offers, the Bayesian analysis indicates that this result should be interpreted with caution because there is no strong support for the effect of experiences of (un)fairness on TPP in children.

Discussion of Study 2A

In Study 2A, older children became increasingly more selective in their enactment of TPP. Also, I found that costly TPP emerges between 5 and 6 years of age, providing converging evidence to prior research (e.g., McAuliffe et al., 2015). These results suggest that children develop a sophisticated understanding of fairness norms and their application in a third-party context over development.

One important question concerned if children's experience as a recipient would influence their willingness to intervene against unfair allocations as a third-party. I found that if anything, receiving a series of unequal offers discouraged children's subsequent third-party punishment. This result would be consistent with the hypothesis that experience of unfairness would decrease costly TPP, which opposes the prediction of the simulation account that the experience of

unfairness would promote TPP. However, further analysis using Bayes factors revealed that our data is moderately in favor of the absence of condition effect, suggesting that the significant main effect of condition found in Study 2A is questionable. Therefore, I conducted Study 2B to replicate the reduced TPP in the unfair experience condition.

Furthermore, in Study 2B, I tried to disentangle two possible causes of the decrease in TPP in the unfair experience condition. One possibility is that children in the unfair experience condition showed reduced TPP because they feel coin-deprived. That is, even though children across three conditions received 20 coins as their initial endowment to play the third-party game, receiving 0 coins for four consecutive trials in the second-party game might have induced children to hold on to the remaining coins, leading to a decrease in costly TPP because of the cost.

The alternative possibility is that during the second-party experience phase, children formed an expectation of how individuals treat each other in this game. That is, when children themselves were treated unfairly by their social partner, this mistreatment might have changed their expectations about how people should treat each other (e.g., "It seems fine to treat each other unfairly in this game as I was treated unfairly"), and thus did not feel compelled to intervene when someone acted unfairly.

Study 2B was designed to disentangle these two possibilities by introducing non-social conditions in which a computer (instead of a peer divider) allocates coins in the second-party game. The number of coins children received in these non-social conditions were matched with those from the fair experience and unfair experience conditions except that children received coins based on the decision made by a computer. Additionally, I assessed how children's

expectation about a divider's offer changes after their own experience of receiving either equal or unequal offers.

Study 2B

Method

Participants. Our final sample were $N = 160$ 5- to 9-year-old children ($M = 89.64$ months, range = 60 - 119 months, $n = 32$ in each age group, 40 per condition, 80 male and 80 female) from the same population as in Study 2A. Demographic information such as race, education, and income was not assessed in this context.

Nine additional children were excluded because they failed to answer correctly in at least one of the comprehension check questions about the consequences and costs involved in each button until the last practice trial (2), the child wanted to stop the study before the test phase (3), there was an experimental error (3) or the parent declined to provide the child's date of birth (1).

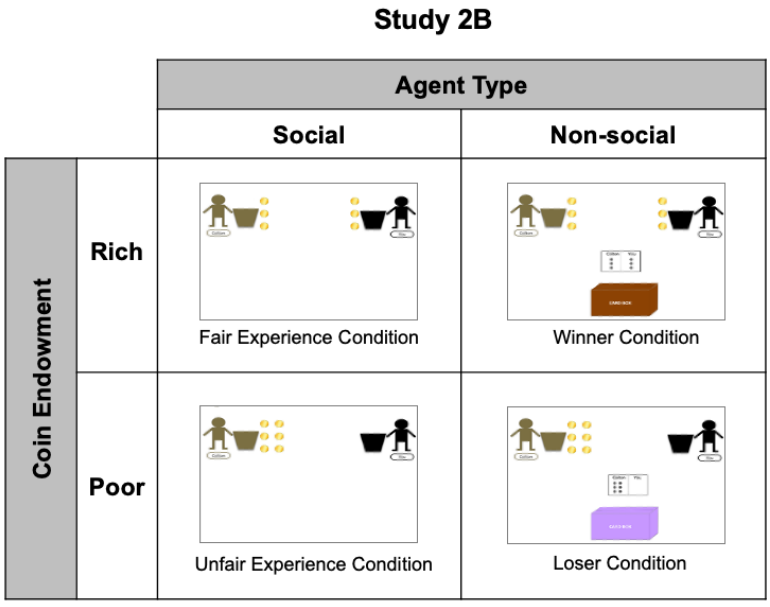

Figure 7. Schematic representation of the design in Study 2B.

The figure illustrates how the second-party game in each condition differed by coin endowment (rich vs. poor) and agent type (social vs. nonsocial). In this example, 6 coins were divided between Colton (left) and the child participant (right). The labels "Fair", "Unfair", "Winner", "Loser", "Rich", "Poor" etc. are added here for representation of the experimental design. Children never heard these words at any point in the study.

Experimental design. A new feature was that there were two additional conditions: *winner* condition and *loser* condition. These two conditions differed from the fair experience condition and unfair experience condition in that the offers that children received during the second-party game were decided by a computer (not by a divider). These conditions were introduced to disentangle the two possibilities described above (i.e., whether children reduced punishment because they felt coin deprived or because they formed a certain expectation about how people treat each other in the game). To address this question, I had a 2 X 2 experimental design (see Figure 7) that differed by coin endowment (whether children received 3 coins every round [rich] vs. whether they received 0 coins every round [poor]) and agent type (whether the offer children received was decided by a social agent [divider] vs. whether the offer was decided by a non-social agent [computer]).

The number of coins that children received during the second-party game was identical in both winner and fair experience conditions. Children in both conditions received 3 coins every round for four consecutive rounds. Similarly, the number of coins that children received during the second-party game was identical in both loser and unfair experience conditions. Children in both conditions received 0 coins every round for four consecutive rounds. 99% of the child participants correctly reported the number of coins they received during the second-party game, confirming our manipulation of coin endowment.

The critical difference between social conditions (fair experience and unfair experience) and non-social conditions (winner and loser) was that children in the non-social conditions were

told that a computer decides how to divide coins between the child participant and another player. Concretely, in non-social conditions, children saw two boxes on the computer screen. In one box, all the cards showed equal offers (the child gets 3 coins and the other player gets 3 coins), whereas all the cards in the other box showed unequal offers (the child gets 0 coins and the other player gets 6 coins). Children were told that the computer will choose one of the boxes and will decide how many coins the child can get. Once the computer chose one of two boxes, the same box was used for drawing a card throughout the rounds during the second-party game. For example, children in the winner condition saw that the computer chose the box with cards of equal offers and drew a card showing an equal offer in every round. Children (98%) correctly identified the agent (another player vs. computer) that is making the offers in the second-party game.

There are at least three possible outcomes. The first possibility is that children in poor conditions (i.e., unfair experience and loser conditions) will show less TPP than those in rich conditions (i.e., fair experience and winner conditions). This finding would imply that children's TPP is driven by the amount of coins they have or the perception that they are coin-deprived.

A second possible outcome is that only those in the unfair experience condition will show a reduced rate of TPP compared to those in the other conditions. More specifically, children in the unfair experience condition might punish less often than those in the loser condition. Such a result would suggest that having an unfair experience from a peer divider would make children assume that selfish allocations are the norm in this game and therefore be less likely to intervene. The result would imply that TPP is influenced by children's expectations about a typical way of dividing resources.

A third possibility is that children's punishment will not differ depending on their experiences of (un)fairness. Thus, rather than an effect of experience condition, I would expect an effect of offer type mediated by an effect of age. Such a finding would suggest that the known developmental effect of children becoming more selective in their punishment as they mature is robust against any contemporaneous experiences. In other words, this potential outcome would indicate that TPP is driven by the development of fairness norms rather than a child's immediate second-party experience of being treated fairly or unfairly.

Procedure. I used the same procedure as in Study 2A, except for the following modifications. First, before and after they played the second-party game, the experimenter asked children to predict how a *new* divider will share six coins with a *new* recipient who did not have any previous history of social interactions or resource allocations. These questions were asked to see whether children's expectations about resource allocations change depending on their second-party experience. Therefore, I posed this question before children had experienced another peer's or a computer's offer in the second-party game and once again after the second-party game. I predicted that those who received allocations made by a computer (i.e., non-social conditions) will not change their expectations about how other people divide coins in this game, whereas those who received offers made by a player (i.e., social conditions) will change their expectations depending on whether they experience fair or unfair sharing.

Second, I did not include the baseline (no experience) condition. Instead, I had a 2 X 2 experimental design as described above.

Third, unlike Study 2A, children played the second-party game *before*, not after, the practice trials of the third-party game. During piloting, I found that children in non-social

conditions had difficulties with the previous order of games used in Study 2A and showed less confusion when they played the second-party game before practicing the third-party game.

Lastly, in this experiment, I did not include the perceived closeness measure because there was no difference in perceived closeness across conditions in Study 2A.

Results

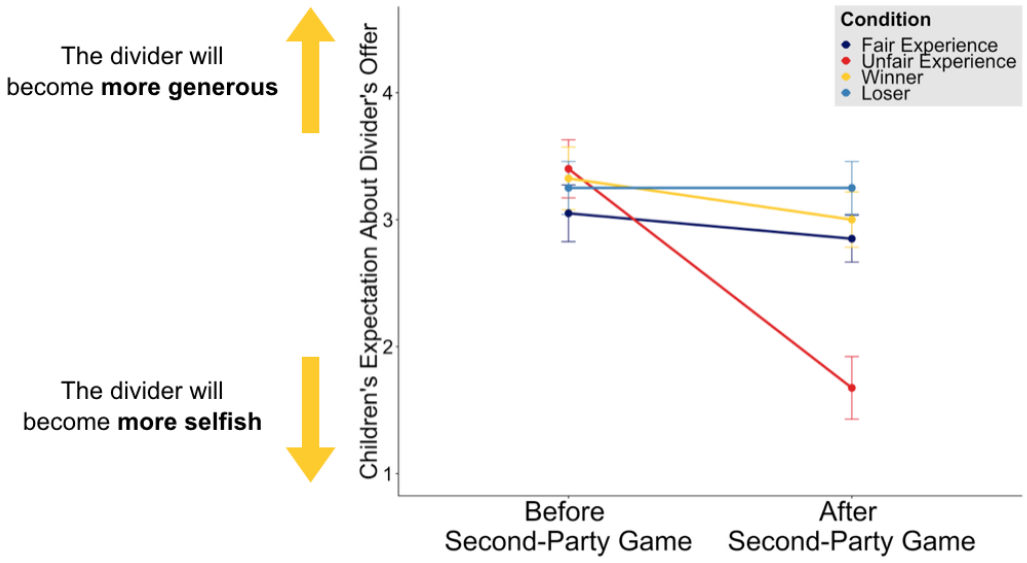

Figure 8. Children's expectation about a divider's offer before and after the second-party game.

In the Y-axis, 3 indicates children's expectation of an equal offer (i.e., sharing 3 out of 6 coins). Error bars represent 95% confidence intervals.

Change in children's expectations about offer. I first calculated the change in children's expectations about the divider's offer by subtracting their prediction *before* the second-party game from that *after* the second-party game. Then, I ran a full linear model on children's change in the expectations about the offer with agent type (social vs. nonsocial), coin endowment (poor vs. rich), age, and interactions among the predictors. The comparison between the full model and the null model with the only intercept indicated that the full model provided a

better fit to the data, $F(7, 150) = 2.99$, $p < .01$[2]. There was no main effect and no interaction effect involving age (all ps > .54). Critically, I found a significant interaction effect between agent type and coin endowment on children's change in expectations, $F(1, 153) = 9.09$, $p < .01$ (see Figure 8).

Specifically, in social conditions, children who received unequal offers in the unfair experience condition ($M = -1.73$, $SD = 2.12$) changed their expectations about the offer more dramatically and expected more selfish offers from another divider compared to those who received equal offers in the fair experience condition ($M = -0.24$, $SD = 1.88$; $F(1, 75) = 10.68$, $b = 1.49$, $SE = 0.46$, $p < .01$). In contrast, in conditions in which allocations were decided by a computer, children's expectations about another divider's offer did not differ between the winner ($M = -0.33$, $SD = 1.82$) and loser conditions ($M = 0.00$, $SD = 1.69$; $F(1, 77) < 1$, $p > .41$).

These results confirmed that agent type was successfully manipulated given that children differentiated offers from social agents vs. non-social agents. Interestingly, those who experienced unequal offers from a social agent generalized their own experience to predict how another divider would treat a third-party. In contrast, those who experienced the same, unequal offers from a computer did not generalize their own experience to predict a divider's offer. That is, children generalized their experience with a social agent to another social agent but not their experience with a non-social agent to a social agent. This result suggests that when forming expectations about how individuals will share coins in this game, children do not rely on mere observations of allocations (i.e., how frequently did the offer occur?), but consider whether the

[2] I excluded two participants' predictions about the divider's offer due to an experimental error during this phase. However, their responses in the third-party game were still included in the analysis of rejection rate as there were no experimental errors during the critical test phase.

allocation was an act of giving by a social agent, in which case they adjust their expectations on how new individuals will behave in this context.

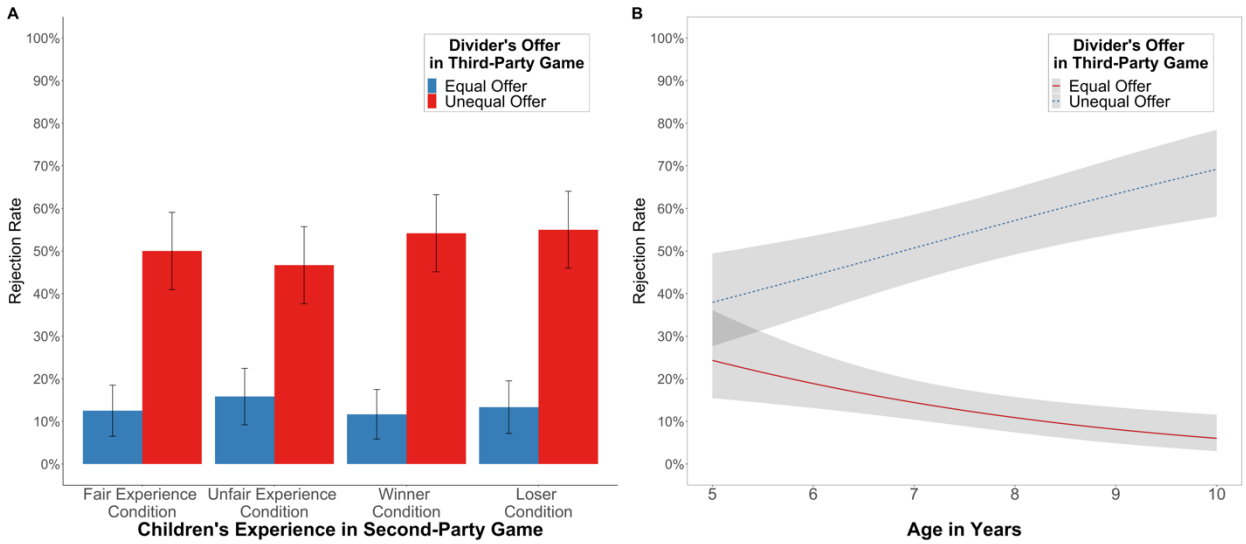

Figure 9. Children's punishment decisions in Study 2B.

(A) Children's rejection rate by condition in Study 2B. (B) Estimates of rejection rate based on the final model in Study 2B (collapsed across conditions). Error bars and confidence bands represent 95% confidence intervals.

Rejection rate. A full GLMM on children's punishment decision (0 = acceptance, 1 = rejection) with agent type (social vs. nonsocial), coin endowment (poor vs. rich), offer type (equal vs. unequal), age, and interactions among the predictors as fixed effects and subject ID as a random effect provided a significantly better fit to the data than the null model with only random intercepts (LRT, χ^2 (15) = 207.51, $p < .0001$).

One critical question was whether children in the unfair experience condition would punish less often than those in the other conditions, which can be determined by an interaction between coin endowment and agent type. The results revealed that there was no two-way

interaction between agent type and coin endowment (LRT, χ^2 (1) = 0.05, p > .82), suggesting that children showed a similar rate of TPP across conditions (see Figure 9).

Another question was whether TPP is influenced by the amount of coins they had or a feeling of coin-deprivation, which can be determined by a main effect of coin endowment. I found that children in poor conditions and rich conditions did not show a significant difference in their rate of TPP (LRT, χ^2 (1) = 0.04, p > .83). Also, children in social conditions and those in non-social conditions did not differ significantly in their rejection rate (LRT, χ^2 (1) = 0.63, p > .42), implying that whether children received offers from a computer or a peer did not lead to a differential TPP. There were no other significant interactions involving either agent type or coin endowment (all ps > .08). These findings suggest that TPP in children is affected by neither the amount of coins they received nor an agent who treated them fairly or unfairly.

The Bayes factors also confirmed the non-significant effects. I found strong evidence for an absence of an interaction between agent type and coin endowment (BF_{01} = 10.18), suggesting that these data are about 10 times more likely to be observed under the hypothesis predicting that there is no such interaction effect than the hypothesis predicting the interaction. Also, the BF of a three-way interaction among agent type, coin endowment, and offer type (BF_{01} = 4.03) revealed that the data is more consistent with the hypothesis predicting that there is no three-way interaction than the hypothesis predicting the interaction effect. BFs for main effects of coin endowment or agent type (BF_{01} = 15.72 and BF_{01} = 10.91, respectively) suggest that the data is more consistent with the hypothesis that punishment was affected by neither the number of coins they had nor the type of the agent who made offers to the child.

As in Study 2A, I found a significant interaction between age and offer type (LRT, χ^2 (1) = 23.62, b = 0.82, SE = 0.17, p < .001). To unpack this interaction, I ran separate models for

equal and unequal offers and found that children were less likely to punish equal offers as they grow older (LRT, χ^2 (1) = 9.00, b = -0.47, SE = 0.16, p < .01), replicating the results from Study 2A. Furthermore, they were more likely to punish unequal offers with age (LRT, χ^2 (1) = 11.55, b = 0.41, SE = 0.12, p < .001). In other words, as they grow older, children's tendency to punish unequal offers increases, while the tendency to punish equal offers decreases. From 66 months of age (5.5 years), the confidence interval of equal offers no longer overlaps with that of unequal offers. I again found very strong evidence in favor of the interaction between age and offer type (BF_{10} = 8290), indicating that the data is 8290 times more likely to be observed under the hypothesis predicting the interaction over the hypothesis predicting no interaction effect. Together, these findings confirm the findings in Study 2A that (1) costly TPP emerges between 5 and 6 years and that (2) children's punishment becomes more selective with age.

Discussion of Study 2B

In Study 2B, children showed a similar rate of costly punishment regardless of their experience as a second-party. It is unlikely that children's TPP is influenced by the amount of coins they received, given that children in poor conditions and rich conditions showed a comparable rate of TPP. Also, our findings are inconsistent with the hypothesis that children's inference on how individuals treat each other in the game affects their TPP. Specifically, children in the unfair experience conditions did not show a significant difference in TPP from those in other conditions. Interestingly, I found that children who received unequal offers from a social agent expected more selfish offers, while those who received unequal offers from a non-social agent did not show a dramatic change in their expectations. Yet, the change in children's expectation about offers in the unfair experience condition did not result in a significant decrease in punishment rate.

One major purpose of this study was to assess the robustness of the initial finding that children punish *less* after an unfair than fair experience. Although the effect found in Study 2A was statistically significant, our Bayes factor analyses had already indicated that this effect might not be strong. In fact, Study 2B showed that this effect did not replicate. Moreover, in both experiments, the Bayes factors for the effect of condition were in favor of the absence of the effect. Based on these results across experiments, it is more reasonable to conclude that there is a lack of support for the influence of second-party experience on TPP.

However, there were findings that have been consistent and strong in both Study 2A and 2B. I found that children become more selective about the enactment of TPP. In Study 2B, with increased age, children were less likely to punish equal offers and were more likely to punish unequal offers, showing a better understanding of when a third-party's intervention is needed. These findings are consistent with the hypothesis that with age, children develop sophisticated general fairness principles.

Discussion of Studies 2A & 2B

The current study examined the influence of experiencing (un)fairness on TPP in 5- to 9-year-old children. This study was possible owing to our novel computer game task in which children's experiences of (un)fairness were manipulated systematically. I started with the hypothesis that being exposed to unfairness personally would enable children to better simulate a third-party's perspective, leading to an increase in following TPP. However, across two experiments, I found no evidence supporting the hypothesis. If anything, in Study 2A, children showed a decrease in TPP after receiving unfair offers. Further analyses revealed that this effect was overall weak and did not replicate in Study 2B. Here, children showed similar rates of TPP regardless of whether they experienced fair or unfair offers. Furthermore, Bayesian analyses

confirmed that the data in both Study 2A and 2B do not support the influence of sharing experiences. Together, these findings suggest that at least in the short-term, immediate experience of (un)fairness does not affect subsequent TPP in children aged 5 to 9.

While there was no indication that prior experience affected children's punishment, there was a very strong and consistent pattern regarding TPP observed in both experiments. I found that with age, children became more selective about the enactment of TPP: They became more likely to consistently punish unequal offers over equal offers. These results support the hypothesis that children develop more sophisticated fairness principles that guide how to divide resources and how to respond to a third-party's unfair act. Moreover, these results show that children punish for fairness reasons and not out of spite or some reason unrelated to fairness. Lastly, Study 2B showed that children punish unfairness regardless of the relative amount of resources they have available. For children, it is not a luxury to intervene against unfairness when they have a surplus of resources, but they punish even when they have been relatively deprived of resources.

Our findings contribute to the existing literature by showing that children employ fairness in an increasingly principled fashion over development. Previous research has suggested that with age, children not only endorse the fairness norms verbally but also adhere to the norms behaviorally (Rizzo & Killen, 2016; Smith et al., 2013). Furthermore, they apply the fairness norms in a consistent manner even when they could be disadvantaged by applying the fairness norms (Blake & McAuliffe, 2011; Shaw & Olson, 2012) and even when their out-group members are disadvantaged (Elenbaas, Rizzo, Cooley, & Killen, 2016; Jordan et al., 2014). Our results are consistent with the development of principled application of fairness norms. From around age 6, children were willing to pay a personal cost to enforce the fairness norms on

another individual even when they were an unaffected third-party. Moreover, this tendency became increasingly selective and systematic with age. Importantly, our data suggest that children's enforcement of fairness norms is unlikely to be swayed by their immediate personal experiences (e.g., How many coins did I get? How did another person treat me?).

One potential concern is that the non-significant effect across experience conditions might be due to a lack of understanding of the task or some confusion about the resources. However, this possibility is unlikely given the ease with which children passed manipulation checks. Specifically, virtually all children (99%) in Study 2A and 2B correctly reported the number of coins they received during the second-party game, suggesting that they paid attention to the information. Also, most children (98%) in Study 2B correctly identified which agent had shared or not shared with them, suggesting that they paid attention to whether the offers were made by a peer or a computer. Furthermore, children expected more selfish offers in the unfair experience condition but not in the loser condition. This result suggests that children were able to differentiate offers made by a social agent from those made by a computer and readjusted their expectations about resource allocations accordingly. Based on the results, it would be reasonable to conclude that the lack of a condition difference is not due to a lack of understanding of the task.

One remaining question is why second-party experience did not influence children's punishment. One possibility is that the second-party experience of (un)fairness was too short and harmless to change TPP. Being exposed to unequal offers briefly may not be sufficient to induce such a change. In fact, studies with adults using more extreme moral norm violations suggest that victims who experienced extreme suffering (e.g., abuse, violence) were more likely to report feelings of empathy for victims who are undergoing suffering and were more likely to participate

in volunteer activities to help victims compared to those who did not experience such sufferings (see Staub & Vollhardt, 2008 for a review). Conversely, it is also known that people who were exposed to abuse or violence are more at risk of being a perpetrator than those who were not exposed to these events (Craig & Sprang, 2007; Dodge, Bates, & Pettit, 1990; Fagan, 2005). In either case, it is possible that more extreme, long-term experiences of unfair treatment could have induced a change in children's TPP. However, in the current study, I was only able to manipulate short-term, mild experiences of unfairness in children but not extreme or long-term unfair experience in children — an experimental manipulation that is not possible for ethical reasons. One approach for future research would be to collect data on children's experiences of unfairness in their daily life and relate it to their willingness to intervene in a third-party context.

Another possibility is that children might not have made a connection between their own experience and the other peer's experience. That is, children failed to take the perspective of the recipient in the third-party game. Prior research with adults has shown that adult participants, who were instructed to take the perspective of the recipient of an unfair offer as much as they can, showed a higher TPP than those instructed to take an objective perspective and remained detached from the recipient (Pfattheicher, Sassenrath, & Keller, 2019). This study implies the importance of taking the victim's perspective in TPP which might not be a skill automatically employed by children. Future research could examine if children would show an increase in TPP when they are given more explicit instruction to take the perspective of the recipient.

Chapter 4: Study 3

Do Children Punish Unfairness to Deter Personal Mistreatment?

Abstract

Third-party punishment has been claimed as a mechanism to deter personal mistreatment. The current study examined if third-party punishment in children is motivated to prevent unfair treatment directed to the self. In this pre-registered study, $N = 120$ 5-to-9-year-olds played a third-party game in which they could decide whether to punish a divider who made unfair resource allocations to another person. Critically, before playing the third-party game, children were told that either the same, selfish divider (Same divider condition) or a new, different divider (Different divider condition) would share resources with them in a game followed by the third-party game. If children punish unfairness to deter future mistreatment, they will enact punishment more often when they need to re-encounter the same divider than when they do not have to. Contrary to the hypothesis, however, I found that children's punishment rate does not change depending on the possibility of future interactions with the same divider.

Do children punish unfairness to deter personal mistreatment?

What motivates children's third-party punishment against unfairness? While little data exists to assess the proximate causes of children's third-party punishment, we can draw inferences from a large body of work with adults. One prevailing theory focuses on the goal to benefit the group. Specifically, this approach views punishment as a mechanism to eradicate selfish individuals to improve the welfare and survival of the group (Fehr & Gächter, 2002; Fehr & Fischbacher, 2003; Fehr & Fischbacher, 2004b; Gurerk et al., 2006). For example, when faced with extinction threats such as wars, famines or environmental catastrophes, groups with individuals who sanction free riders would be more likely to survive than groups with no punishment towards free riders (Fehr, Fischbacher, & Gächter, 2002; Mathew & Boyd, 2011). This approach assumes that people have an altruistic disposition to punish selfish behaviors (Fehr, Fischbacher, & Gächter, 2002; Fehr & Fischbacher, 2004b).

On the contrary, another common theory claims that punishment serves future benefits for punishers themselves, not their group. To elaborate, this approach contends that adults punish perpetrators to deter personal exploitation and to get a better bargain for themselves directly (Peterson, Sell, Tooby, & Cosmides, 2010). Studies have shown that people are more likely to punish unfair dividers when they inferred that the divider would treat them poorly than when they inferred that the divider would treat a third party poorly (Delton & Krasnow, 2017; Krasnow, Delton, Cosmides, & Tooby, 2016). For instance, the more people predicted that they would be mistreated by the divider, the more third-party punishment they enacted towards the divider. Interestingly, punishers' prediction about how the divider would treat a third-party was not correlated with the enactment of third-party punishment (Krasnow et al., 2016). These

findings suggest that a concern that punishers themselves would be mistreated by the divider (not a concern for a third-party victim) underlies third-party punishment in adults.

To our knowledge, there is no research that has directly tested whether children's third-party punishment is motivated by potential possibility of personal mistreatment. More specifically, here I examine if children's punishment rate changes depending on whether they will interact with the same, selfish divider in the future. I start with the notion that at least in some contexts, children start to think about future from age 3 and engage in more sophisticated prospective thinking from age 5 (Atance & Meltzoff, 2005). Therefore, it is possible that children's emerging ability to think about the future enables them to punish unfair dividers to deter potential mistreatment of the self. The current study investigates whether children enact third-party punishment to deter personal mistreatment.

To examine this question, I presented participants with a third-party game in which they observed either fair (3:3) or unfair allocations (6:0) between a divider and a recipient as an unaffected third-party. Critically, before playing the third-party game, child participants learned that in the subsequent two-party game, they will receive resources (e.g., coins) from the same divider from the third-party game (*Same divider condition*) or from a new divider (*Different divider condition*). Thus, depending on condition they were assigned, children learned whether they need to interact with the same divider again in the future or not.

I tested the hypothesis that children would use punishment to deter potential mistreatment in the future ("Deterrence hypothesis"). Concretely, if children use third-party punishment as a way to prevent future mistreatment, they will punish unfair allocations *more often* during the third-party game when they have to interact with the same divider in the following game (Same divider condition) than when they do not have to interact with the same divider (Different divider

condition). Alternatively, if children fear retaliation from the divider ("Appeasement hypothesis"), they will punish unfair allocations *less often* when they have to interact with the same divider in the future (Same divider condition) compared to when they do not have to interact with the same individual (Different divider condition). Our study was designed to assess these different hypotheses.

Method

Participants. Our final sample were $N = 120$ 5- to 9-year-old children ($M = 89.2$ months, range = 61 - 120 months, $n = 60$ per condition, $n = 24$ in each age group, 60 male, 60 female). Children were tested at a museum or public parks in the Midwest of the US. Demographic information such as race, education and income could not be obtained. Thirteen additional children were excluded because of failure to correctly answer at least one of the comprehension checks (9), parental interference (2), being unable to understand English (1) or having participated in a similar study involving third-party punishment before (1).

Experimental design and procedure. After parents gave written consent, children sat at a table with the study apparatus while the parents watched passively from a few steps away. A female experimenter introduced the computer game referred to as the "coin game" and explained that players could collect virtual coins to later exchange for prizes. During a *prize introduction*, children learned that the more coins they have during the coin game, the more and the better prizes they would be able to choose afterwards.

In the subsequent *practice phase*, the experimenter introduced the two other players in the game by stating that they were children of the same age and gender at another museum (or another park), who are currently connected online. The experimenter introduced children to a third-party game (see Figure 10) in which they can decide whether to punish an allocation as an

uninvolved third-party observer. Specifically, children learned the role of the divider and the recipient. The divider could make one of two allocations: (a) 3 for the self and 3 for the recipient, or (b) 6 for the self and 0 for the recipient. The recipient was a passive player who could only accept the divider's allocation. During the practice of the third-party game, children learned the consequences of pressing the green (acceptance at no costs) or red button (rejection at a cost).

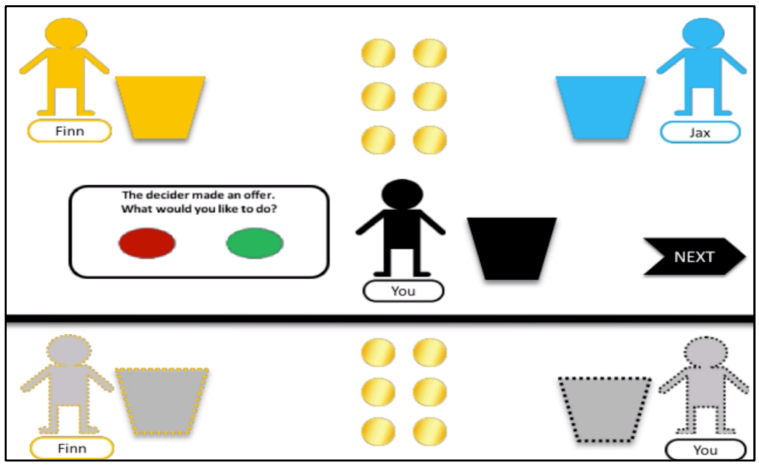

Figure 10. Computer screen display of the coin game in the same divider condition. In this example, Finn was the divider in both third-party and second-party games.

In the subsequent *manipulation phase*, children in both conditions were told that, after playing the third-party game, they will play a second-party game in which a divider allocates coins between the self and the child participant. Critically, children were assigned to one of two conditions. In the *Same divider condition*, children were told that the same divider from the third-party game will decide how to share coins with the child. In the *Different divider condition*, they were told that a new divider will decide how to share coins with the child.

In both conditions, to visualize the possibility of future interactions with the same or different divider, the computer game display indicated the future divider in the second-party

game (see the bottom of the display in Figure 10). For example, in the same divider condition, the experimenter told children that the same divider from the third-party game (e.g., Finn in Figure 10) will be moved to the bottom left corner on the screen in the second-party game, and the child participant will be moved to the bottom right corner on the screen. In the different divider condition, children learned that a new divider (e.g., Jax in Figure 11) who does not play the third-party game at all will appear at the bottom left corner and the child participant will be moved to the bottom right corner in the second-party game.

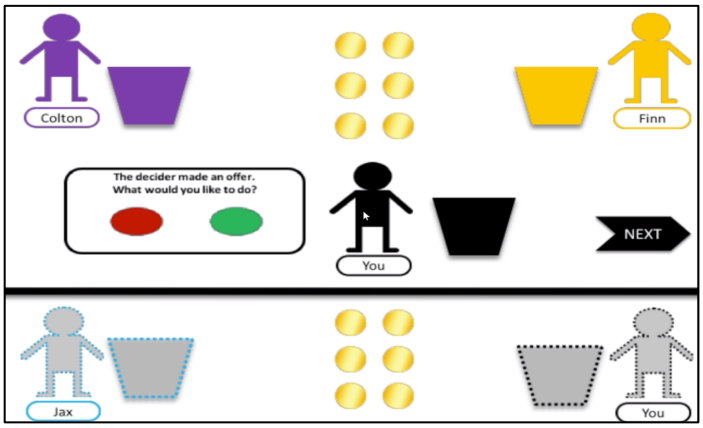

Figure 11. Computer screen display of the coin game in the different divider condition. In this example, Colton was a divider in a third-party game, while Jax was a new divider in the second-party game.

In both conditions, the experimenter told children that both the divider and the recipient in the third-party game could see the child participant's decisions to reject or accept an allocation. Children in the different divider condition were additionally told that the new divider could not see their decisions to reject or accept during the third-party game.

At the end of the manipulation phase, the experimenter asked comprehension check questions to assess (a) whether children could correctly identify the divider and recipient in the following second-party game and (b) whether they understood that the divider and recipient in

the third-party game (but not a new divider in the second-party game) could see their punishment decisions. I found that most children were able to correctly identify the divider (83% of participants) and the recipient (93% of participants) in the second-party game. Also, most children understood that the divider and the recipient in the third-party game could see their punishment decisions (91% and 88%, respectively, across two conditions), while the new divider could not (87% from those in different divider condition). When children answered incorrectly in these comprehension check questions, the experimenter provided a correct answer to the child.

After learning about the possibility of re-encountering the same divider in the future, children in both conditions played the identical third-party game (*test phase*) in which they can decide whether to punish an allocation by pressing either green (acceptance) or red button (rejection). Children were presented with 8 allocation trials in total. Importantly, the divider made an unfair allocation (6:0) in 6 trials, while he or she made a fair allocation (3:3) in 2 trials. This manipulation was intended to make children perceive the divider as a selfish individual, which should in turn allow children to predict that the divider would share resources in a selfish manner with the child participant in the subsequent second-party game.

Upon the completion of the third-party game, to check whether children perceive the divider from the third-party game as a selfish individual, the experimenter asked children to recall whether the divider kept all coins for the self most of the time or whether the person shared coins with the recipient most of the time. The result was that a vast majority of children (83%) correctly recalled that the divider in the third-party game kept all coins for the self in a majority of the trials. In addition, children were asked to predict whether the divider in the second-party game would share fairly (3 out of 6 coins) or unfairly (0 out of 6 coins) with the child participant. I predicted that children in the same divider condition should be more likely to predict unfair

sharing from the second-party game's divider (as the same divider had been selfish during the third-party game) than children in the different divider condition who were asked to make a prediction about a new divider who did not have any sharing history. Contrary to the prediction, in both conditions, I found that a majority of children (77%) expected a fair sharing (i.e., receiving 3 out of 6) from the same divider as well as different divider (46 out of 60 participants expected fair sharing in both conditions). This result suggests that children have an optimistic expectation that they will be treated fairly even in the same divider condition, in which their own observation (i.e., the divider was selfish towards a third party in 6 out 8 trials) directly contradicted their optimistic expectation. Also, this optimistic expectation cannot be attributed to children's difficulty with remembering the divider's behavior because most children correctly recalled that the divider in the third-party game acted selfishly in most of the time.

At the end of the study, the experimenter left and a secondary experimenter asked children whether they thought the players were real or pretend. I found that 80% of children said the players were real.

I counterbalanced the order of test trials, practice trials, comprehension check questions, position of green and red buttons, and the other player's identity.

Data coding and analyses. Children's responses were automatically recorded by GameMaker Studio (https://www.yoyogames.com) and later checked and entered into a spreadsheet by independent coders. All statistical analyses were conducted with R statistical software (R version 3.5.2; R Core Team, 2018).

I compared a full Generalized Linear Mixed Models (GLMM), which included age in months and condition and an interaction between age and condition as fixed effects and subject ID as a random effect with a null model, which included only subject ID as a random intercept. If

the full model provided a significantly better fit to the data, I created a minimal model by sequentially dropping single terms from the full model, and finalized our minimal model when dropping single terms no longer provided a better fit to the data.

Results

Preliminary analysis. I first analyzed whether children were more likely to reject unfair over fair allocations. A full GLMM on children's punishment decision (0 = acceptance, 1 = rejection) with offer type (fair vs. unfair), age and an interaction between offer type and age as fixed effects and subject ID as a random effect provided a significantly better fit to the data than the null model with only random intercepts (LRT, χ^2 (3) = 82.79, $p < .001$). I found a significant main effect of offer type on punishment decision (see Figure 12A), LRT, χ^2 (1) = 82.64, $p < .001$. These results confirm that children punished unfair allocations more often than fair allocations as shown in other studies with children (e.g., House et al., 2020; Jordan et al., 2014; McAuliffe et al., 2015). There were a non-significant interaction between offer type and age, LRT, χ^2 (1) < 1, $p > .70$, and a non-significant main effect of age, LRT, χ^2 (1) < 1, $p > .91$.

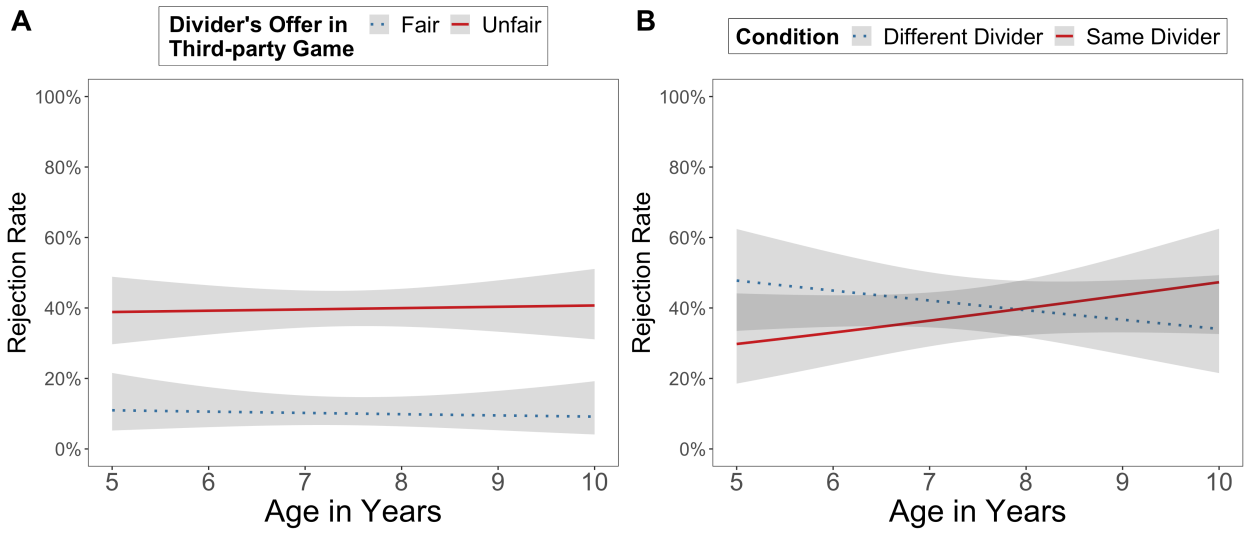

Figure 12. Children's punishment decisions in Study 3.

(A) Estimates of rejection rate by offer type and age. (B) Estimates of rejection rate by condition and age.

Rejection rate. As there were more unfair trials than fair trials (6 vs. 2 trials), in the further analysis, I excluded children's responses to fair allocation trials and analyzed responses to unfair allocations exclusively. A full GLMM on children's punishment decision (0 = acceptance, 1 = rejection) with condition (same vs. different divider), age and an interaction between condition and age as fixed effects and subject ID as a random effect did not provide a better fit to the data than the null model with only random intercepts (LRT, χ^2 (3) = 3.27, $p >$ .35). These findings indicate that children's punishment of unfair allocations did not differ significantly depending on condition (see Figure 12B).

Discussion of Study 3

The current study examined the deterrence hypothesis in which 5- to 9-year-old children are more likely to punish an unfair individual when they need to interact with the person in the future than when they do not need to. Our results suggest that children overall punish unfair

allocations more often than fair allocations. Critically, however, Children's rate of punishment towards unfair allocations did not differ depending on the possibility of re-encountering the same divider in the future. This finding is inconsistent with the deterrent hypothesis in which children should increase their punishment when they were told that the same divider would decide how to allocate resources between the self and the child participant in the future. Also, our findings do not support the appeasement hypothesis in which children would fear retaliation from the divider and would show a decrease in punishment when told to re-encounter the same divider in the future.

Why did third-party punishment not change depending on condition? One possibility is that children in the current study did not infer unfair treatment of themselves from observing how the selfish divider treated a third-party. This possibility is supported by children's optimistic expectation about fair sharing. Concretely, in the present study, a majority of children expected that they would be treated fairly by a divider regardless of conditions. Interestingly, even children in the same divider condition who witnessed that the divider tended to make a selfish allocation expected that they would be given special treatment and receive a fair offer from the selfish divider in the following game. Hence, it is likely that children's general optimism about how they would be treated led to a non-difference in punishment rate between conditions.

To address children's optimistic bias about self, future research should investigate whether children infer deterrence-based motive when they hear a story about a third-party punisher. If children successfully infer that punishment could be driven by a motivation to deter future mistreatment, it would suggest that children could reason about the function of punishment in terms of deterrence when they are completely uninvolved with the situation, but they are not yet able to incorporate this reasoning into their own punishment behavior potentially

due to their optimistic self-serving bias when they themselves are the recipient. Alternatively, if children do not reason about punishment in terms of deterrence motive even in a context in which they passively observe third-party punishment, it would suggest that at least children around this age do not think of third-party punishment in terms of deterrence motive at all.

Chapter 5: Study 4

Children's Evaluations of Third-party Responses to Unfairness: Children Prefer Helping over Punishment

Abstract

Third-party punishment of selfish individuals is an important mechanism to intervene against unfairness. However, there is another way in which third parties can intervene. Rather than focusing on the unfair individual, third parties can choose to help those who were treated unfairly by reducing inequality. Such third-party helping as an alternative to third-party punishment has received little attention in studies with children. Across four studies, I examined the evaluations of third-party punishment versus third-party helping in $N = 322$ 5- to 9-year-old children. Study 4A, 4C and 4D showed that when asked about the agents directly, children evaluated both helpers and punishers positively, but they preferred helpers over punishers overall. When asked about the type of intervention itself, children preferred helping over punishment, suggesting that their preference for the type of intervention corresponds to how children think about the agents performing these interventions. Study 4B showed that children's preference for third-party helping is driven by distributive justice concerns and not a mere preference for giving or resource maximization as children consider which type of third-party intervention decreases inequality. Together, this series of studies demonstrate that children between 5 and 9 years of age develop a sophisticated understanding of punishment and helping as two adequate forms of intervention but also display a preference for third-party helping. I discuss how these findings

and prior work with adults supports the hypothesis of developmental continuity, showing that a preference for helping over punishment is deeply rooted in ontogeny.

Children's evaluations of third-party responses to unfairness: Children prefer helping over punishment

Notwithstanding the important insights gained from studies of third-party punishment, it is by no means the only way to intervene against moral transgressions. For example, rather than punishing the agent of the act, a third party might choose to help the patient who has been placed at a disadvantage. With punishment targeting the perpetrator and helping targeting the recipient of unfair treatment, third-party punishment and helping focus on different aspects of justice restoration. While third-party punishment has been studied extensively with adults and to some extent in children, very little research has been devoted to third-party helping, despite its equal importance.

Several experiments have tested whether adults want to help or punish when they are the third party. Some studies find that adults would rather like to help the recipient than to punish the perpetrator (Chavez & Bicchieri, 2013; Jordan, Hoffman, Bloom, & Rand, 2016), although others found the reverse preference (e.g., FeldmanHall et al., 2014; Stallen et al., 2018). Findings are more consistent with how adults evaluate punishers and helpers. Overall, adults prefer third parties who help recipients over those who punish transgressors. For example, when observing how one person was unfair to a recipient, adults were more likely to reward a third-party helper who gave resources to the recipient rather than a third party who punished the unfair divider (Raihani & Bshary, 2015a). Similarly, further studies showed that helpers were perceived to have superior moral character (e.g., warmth) compared to punishers (Patil, Dhaliwal, & Cushman, 2018) and are more likely to be trusted as a partner in economic interactions (Jordan et al., 2016; Patil, Dhaliwal, & Cushman, 2018). These results imply that third-party helping is not only

regarded as a viable, important response to unfairness but also gains an even more positive reputation than punishment.

Current Study

Here I aim to examine (1) whether children evaluate third-party punishers positively or negatively and (2) how punishers are evaluated in comparison with helpers focusing specifically on fairness violations. To our knowledge, it has not been studied how children evaluate punishers versus helpers when they are presented as two options to respond to unfairness. Given the finding that adults prefer helpers over punishers, it is important to investigate how children think about these forms of intervention and to trace its developmental trajectory.

One hypothesis is that children would show a pattern of results similar to adults. That is, regardless of their age, children might evaluate both helpers and punishers positively overall, but when forced to choose, they might regard helpers as even more positive than punishers. Under this hypothesis, the preference of adults for helpers over punishers is a psychological phenomenon that already appears early in development when children begin to reason about third-party punishment.

A plausible alternative hypothesis is that children undergo major developmental changes in their preferences, where young children evaluate only helpers positively, while they view punishers as negative or neutral at best potentially due to punishment being an antagonistic behavior (i.e., taking resources away from a person). With increasing age, children might evaluate punishers more positively as they gain a better understanding of potential benefits of punishment such as maintenance of norms and deterrence of transgressions (see Bregant, Shaw, & Kinzler, 2016 for children's understanding of the deterrent effect of punishment).

Here I present a series of studies comprising a pilot study to validate our new task and four studies to address the above hypotheses.

General Methods

Experimental design and procedure. Across four studies, children heard a story about four actors — one divider, one recipient and two third parties — who played a candy game at a summer camp. The experimenter walked children through the story with visual aids by using paper-cut actors and candies. The actors were matched to the gender of the participant.

At the beginning of the story, the divider and the recipient have two candies each. Then, the divider can decide how to divide the additional two candies (see Figure 13). Children were told that the divider can keep both candies or give some of the candies to the recipient.
In fact, because I was interested in how children respond to fairness violations, the divider always kept both candies for themselves, resulting in unequal allocation (4:2 between the divider and recipient).

Before showing children an unfair allocation made by the divider, I assessed whether children endorse sharing by asking what number of candies the divider should give to the recipient. It was important to confirm that children endorse sharing because otherwise, they would not be able to understand the purpose and intention of third-party intervention. Across four studies, a majority of children (80%) endorsed equal sharing (i.e., sharing 1 out of 2 candies) (65 out of 80 participants in Study 4A, 62 out of 80 participants in Study 4B, 65 out of 80 participants in Study 4C, and 66 out of 82 participants in Study 4D), and the remaining 20% stated the divider should give 2 out of 2 candies to the recipient, $\chi^2 (1) = 116.9$, $p < .001$. I excluded children from analyses if they considered it acceptable for the divider not to share with the recipient (i.e., giving 0 candies).

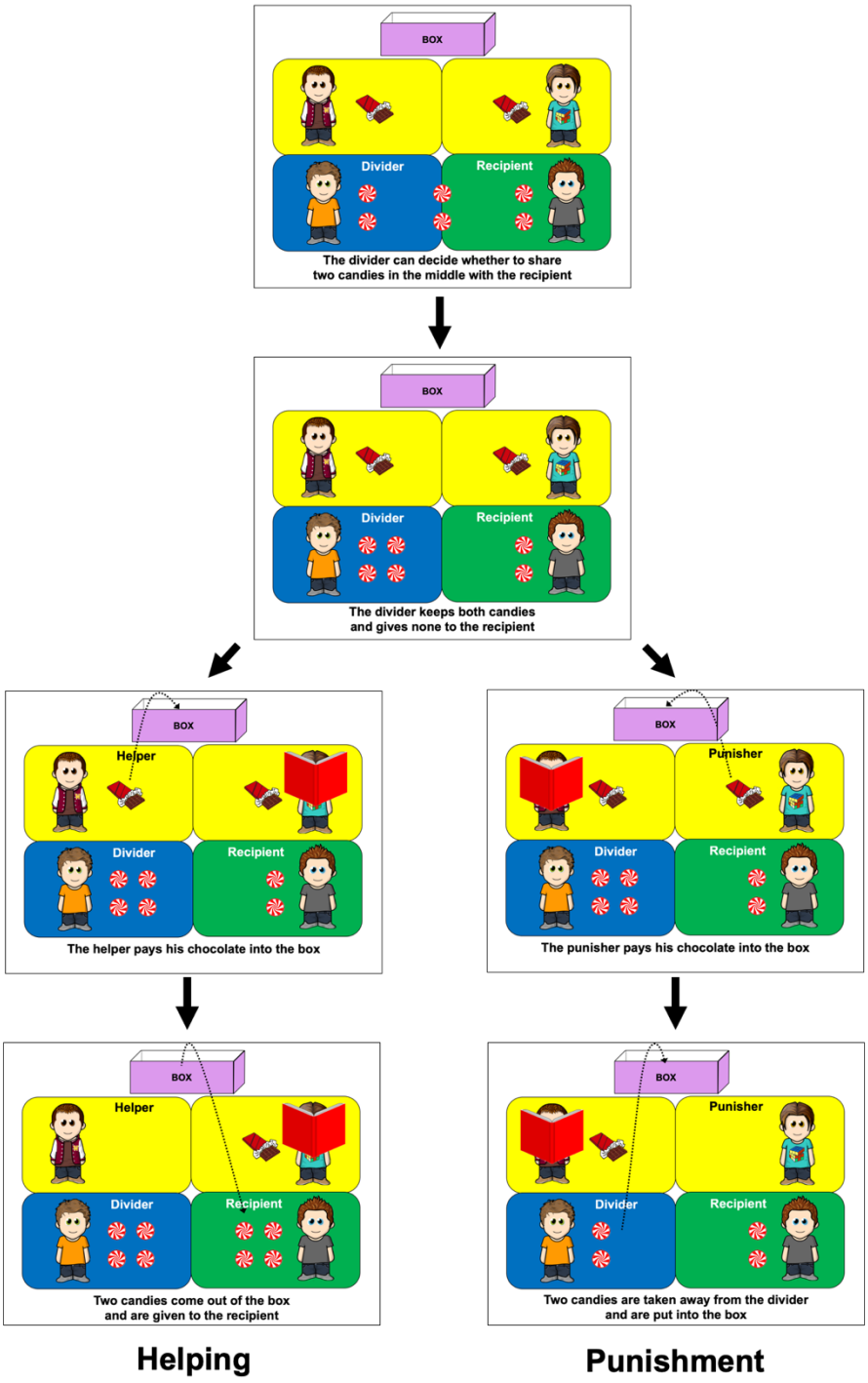

Figure 13. Schematic set-up of stimuli used in Study 4A.

The two panels at the top show an unfair allocation. The two panels on the left show a helping scenario, while those on the right show a punishment scenario. Each panel shows the laminated figures and objects used by the experimenter to illustrate the different scenarios to the child. The labels “Helper”, “Punisher” etc. are added here for reading comprehension. Children never heard the word “helper” or “punisher” at any point in the story.

Table 1. Experimental design with resource allocation (Divider:Recipient)

Study	Condition	Scenario	Before Intervention	After Intervention	Inequality Between Divider and Recipient
Study 4A, 4C & 4D		Helping	4:2	4:4	0
		Punishment		2:2	0
Study 4B	Helping Only	Rational	4:2	4:4	0
		Irrational		6:2	4
	Punishment Only	Rational	4:2	2:2	0
		Irrational		4:0	4

After seeing the unfair allocation, each child heard both punishment and helping scenarios. In each scenario, the third parties have three options: (1) give two candies to the recipient by paying own chocolate (i.e., costly helping), (2) take two candies away from the unfair divider by paying own chocolate (i.e., costly punishment) or (3) do nothing. In the *helping scenario*, one of the two third parties decides to give two candies to the recipient by paying his or her own chocolate (see Figure 13 and Table 1). By contrast, in the *punishment scenario*, the other third-party decides to take two candies away from the unfair divider by paying a chocolate. The payment of a chocolate in this case represents a personal cost that the third-party has to bear. I made each third-party intervention costly because the payment of a cost could remove the doubt in observers that the third-party's intervention is motivated by a self-serving desire and can signal that the third-party genuinely cares about others (Balliet et al., 2011; Nelissen, 2008; Raihani & Bshary, 2015b). In each scenario, only one of the two third-parties was involved. Children were told that the other third-party who was not involved in a given scenario was

reading a book. This was to prevent children from blaming the non-involved third-party for not intervening.

After hearing both scenarios, children were asked test questions. Two primary measures were kept constant across four studies: (1) Children's liking of each third-party actor on a 7-point smiley face Likert scale and (2) their forced-choice preference between the two third parties (e.g., who do you like better?).

During the warm-up phase (i.e., before children heard the story), children practiced how to use a Likert scale to indicate their liking (Study 4A through 4D) and their agreement about an action (Study 4C and 4D). Children were asked to point to one of the faces on the Likert scale for something they really like or really do not like, ranging from -3 (do not like it at all) to +3 (like it a lot). Also, they were asked to point to a face on the Likert scale for something they strongly agree or strongly disagree, ranging from -3 (totally disagree) to +3 (totally agree). Children's responses during the practice confirmed that they can use the Likert scale to indicate their agreement in a flexible manner.

I define third-party helping as a third-party's compensation of payoffs to another individual. This term was used for consistency with other similar studies with adults (e.g., Raihani & Bshary, 2015a; Jordan et al., 2015).

In all four studies, I counterbalanced the order of scenarios (punishment scenario first vs. helping scenario first), the order of the Likert scale (introduce positive rating first vs. negative rating first), the order of comprehension check questions, the order of test questions and appearances of third-party actors.

Data coding and analyses. Children's responses were live coded by the experimenter and later checked by an independent video coder. Disagreements between live and video coding

were resolved by re-watching the video. All statistical analyses were conducted with R statistical software (R version 3.5.2; R Core Team, 2018). In our analyses, I entered the age in months, not in years.

In all studies, I analyzed children's endorsement of equal sharing with a goodness-of-fit chi-square test, their responses in the Likert scale rating with one sample t-tests (two-tailed) and their responses in the forced-choice measures with binomial tests (two-tailed). To test the effect of age, I ran Linear Mixed Models (LMM) on the Likert scale ratings using the package 'nlme' (Pinheiro et al., 2018), Generalized Linear Models (GLM) on the social preference scores and Generalized Linear Mixed Models (GLMM) on forced choices using the package 'lme4' (Bates, Maechler, Bolker, & Walker, 2015).

Our analysis procedure was as follows: (1) I examined a null model, which included only subject ID in mixed models and only intercept in the linear models; (2) I created a full model, which included our main predictors (e.g., intervention type, age in months) and all interactions among the predictors; (3) I compared the full model with the null model; (4) if the full model provided a significantly better fit to the data, I created a minimal model by sequentially dropping single terms from the full model, testing whether their inclusion improved the model fit; (5) I stopped this process and finalized our minimal model when dropping single terms no longer provided a better fit to the data.

I proceeded to test individual predictor variables only if the full model with predictors provided a significantly better fit to the data than the null model. Moreover, since the current four studies comprised a large number of analyses, for the sake of brevity and clarity, I report test statistics only for the final model.

Figures show predicted estimates with 95% confidence intervals based on the final model. All data and protocols are available through the Open Science Framework: https://osf.io/8wxtj/?view_only=f46dc44bf88845cbb93016a91371d62d.

Pilot Study

In our pilot study, I tested $N = 32$ 6- to 7-year-old children. I chose this initial age because prior work had established that children 6 years and older reliably detect violations of fairness norms and are willing to punish transgressions (e.g., Jordan et al., 2014; McAuliffe et al., 2015). I piloted our aforementioned task, where children heard about a third-party who punished an unfair divider, whereas another third-party helped the recipient of the unfair allocation. Our dependent measures were children's liking of each third-party character measured on a 7-point Likert scale and a forced-choice task to choose between the punisher or the helpers. This pilot established that children were able to follow the story, the task, and the measures, as established through comprehension checks. Moreover, the results indicated that children evaluated both third parties positively, but they tended to prefer helpers over punishers. After this first validation, I made minor modifications to streamline the procedure and used this method for a series of four studies. For these studies, I broadened the age range to 5- to 9-year-olds to examine potential developmental change.

Study 4A: How Do Children Evaluate Helpers and Punishers?

Method

Participants. Our final sample were $N = 80$ 5- to 9-year-old children ($M = 89.33$ months, range = 60 - 118 months, $n = 16$ participants in each age group, 40 female). Children were recruited and tested in a public park in an urban area in the US. After parental consent, children were brought to a table with the study apparatus while the parents watched from a few steps

away, instructed not to intervene. Demographic information such as race, education, and income were not assessed in this context. Fourteen additional children were excluded because they failed to identify the helper and punisher during both memory check questions (9) or they did not endorse sharing and said that the divider should give zero candies to the recipient (5).

Experimental design and procedure. I used the same basic design and procedure described in General Methods, with the following additional measures. In the forced-choice questions, I asked not only who they like better between the helper and the punisher (labeled "like better") but also who they want to be friends with (labeled "friend"), who they want to invite to their party (labeled "party").

Furthermore, to explore children's reasoning behind their preference for punishers and helpers, I tested whether children attribute relevant traits to either the punisher or helper. For instance, children were asked who is more likely to get into a fight (to see if they attribute "aggression" to punishers), who is more likely to tell a person who cut in line to go back (to see if they attribute "norm enforcement" to punishers), who is more likely to give more turns to ride a bike to others (to see if they attribute "generosity" to helpers), and who is more likely to comfort a crying child (to see if they attribute "empathy" to helpers). These forced-choice trait attribution questions were asked with other social preference questions (e.g., like better, friend, party) during the test phase. Our trait attribution tasks measured a relative trait inference (e.g., "helper is more generous than punisher") instead of whether children reasoned about a single person (e.g., "helper is generous") (see Liu & Vanderbilt, 2013).

Results

Evaluations in Likert scale ratings. Children evaluated both the helper and the punisher positively. Children's evaluation of both the helper ($M = 2.15$, $SD = 1.29$) and punisher ($M =$

1.84, *SD* = 1.35) differed significantly from neutral (one sample t-tests; $t(79) = 14.87$, $p < .001$ and $t(79) = 12.14$, $p < .001$, respectively). The results from an LMM indicated that neither age nor intervention type (helper vs. punisher) had a significant effect on children's ratings. These analyses show that children's ratings of helpers and that of punishers did not differ significantly at all ages (see Figure 14A). The result from a paired t-test revealed that children tended to rate the helper more positively than the punisher with a marginal significance, $t(79) = 1.72$, $p = .09$.

Social preference in forced-choice questions. I first analyzed all three preference questions separately. Children chose helpers over punishers when asked who they liked better (73%; binomial test, $p < .001$) and who they wanted to invite to their party (66%, $p < .01$). However, their preference for helpers did not differ from chance when asked who they would choose as a friend (60%; $p = .09$; see Table 2). One question is why children's preference was more pronounced in the 'like better' question than in 'party' or 'friend' questions (note that the preferences in these two questions were trending towards helpers as well). I speculate that the friend and party questions might involve other social motivations. For example, regardless of who they like personally, some children might want to associate with the punisher who might seem socially dominant or seem to have power and assertiveness. By contrast, the 'like better' question is a more direct and pure measure of social preference per se.

Children's responses in the three forced-choice questions (friend, party and like better) were intercorrelated (Cronbach's $\alpha = .63$). Thus, in subsequent analyses, I averaged responses to the three forced-choice questions for each child, which I refer to as the *social preference score* (helper = 1, punisher = 0) with higher values reflecting a preference for helpers over punishers. The results from a full GLM including age as a fixed effect revealed that there was no significant effect of age on children's social preference score, $\chi^2(1) = 2.02$, $p > .15$ (see Figure 14C).

Overall, in the forced-choice preference questions, children preferred helpers to punishers irrespective of their age.

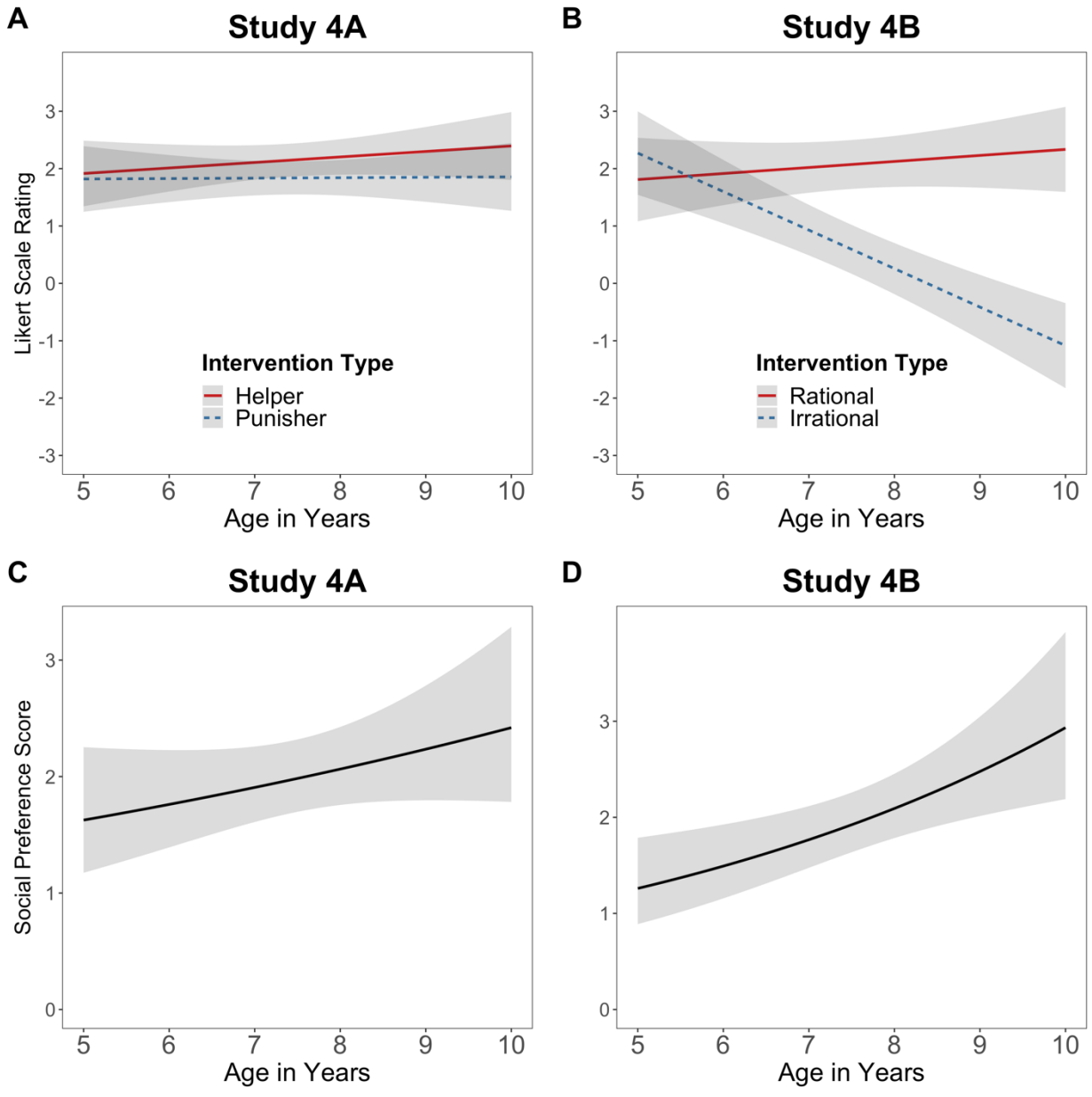

Figure 14. Likert scale ratings and social preference scores in Study 4A and 4B.

(A) Likert scale ratings of helpers and punishers in Study 4A. (B) Likert scale ratings of rational and irrational third parties in Study 4B. (C) Higher score indicates children's preference for

helpers over punishers in Study 4A. (D) Higher score indicates children's preference for rational over irrational third parties in Study 4B.

Table 2. Mean proportion of children who chose helpers over punishers (Study 4A) and those who chose rational over irrational third parties (Study 4B) in binomial tests collapsed across age.

Forced-Choice Question Type		Study 4A	Study 4B	
			Helping Only	Punishment Only
Social Preference	Friend	0.60 [0.48, 0.71]	0.65 [0.48, 0.79]	0.60 [0.43, 0.75]
	Party	0.66** [0.55, 0.76]	0.80*** [0.64, 0.91]	0.48 [0.32, 0.64]
	Like Better	0.73*** [0.61, 0.82]	0.72** [0.55, 0.85]	0.70* [0.53, 0.83]
Trait Attribution	Aggression	0.27*** [0.18, 0.39]	0.20*** [0.09, 0.36]	0.50 [0.34, 0.66]
	Norm Enforcement	0.35** [0.25, 0.46]	0.40 [0.25, 0.57]	0.55 [0.38, 0.71]
	Generosity	0.73*** [0.61, 0.82]	0.75** [0.59, 0.87]	0.55 [0.38, 0.71]
	Empathy	0.68** [0.56, 0.78]	0.80*** [0.64, 0.91]	0.65 [0.48, 0.79]

Note. The numbers in parentheses represent 95% confidence intervals.
* $p < .05$, ** $p < .01$, *** $p < .001$

Trait attribution. Children attributed relevant traits to the third parties: Children attributed generosity (73%; binomial test, $p < .001$) and empathy (68%; $p = .002$) to helpers, while they attributed aggression (73%; $p < .001$) and norm enforcement (65%; $p < .01$) to

punishers (see Table 2). These results suggest that children attributed warmth-related traits (generosity, empathy) to helpers and conflict-related traits (aggression, norm enforcement) to punishers, providing converging evidence with their preference for helpers over punishers.

Discussion of Study 4A

Study 4A showed that children evaluate both third parties positively and their ratings did not differ from each other significantly on the Likert scale. In forced-choice tasks, however, children preferred helpers over punishers. There were no age-related changes in either the Likert scale ratings or forced-choice social preference questions. Children inferred that helpers were more likely to show warmth-related traits (generosity, empathy), while punishers were more likely to show conflict-related traits (aggression, norm enforcement).

However, there are at least two alternative explanations for children's preference for helpers over punishers. One possible explanation is that children prefer helpers not because they reason about responses to unfair treatment but merely because they prefer givers over takers. They might have associated giving with positive evaluation and associated taking with negative evaluation. To address this alternative hypothesis, it is important to test if children like givers regardless of whether justice is restored or not. Another alternative account that could explain our previous results is that children might have preferred those who maximize resources. Concretely, in our setup, helping always resulted in more resources for the actors than punishment. For instance, helping resulted in 4:4 (8 candies in total), while punishment leads to 2:2 (4 candies in total). Under this hypothesis, children do not care about rectifying inequality, but simply think that increasing rewards is generally better. Study 4B was designed to address these possibilities.

Study 4B: Do Children Merely Prefer Givers or Resource Maximizers Without Considering Fairness?

Method

Participants. Our final sample were $N = 80$ 5- to 9-year-old children. One half of the children were assigned to the helping-only condition and the other half to the punishment-only condition. In the helping-only condition, our sample were $n = 40$ ($M = 89.80$ months, range = 60 - 116 months, 8 participants in each age group, 20 female). In the punishment-only condition, our sample were $n = 40$ ($M = 88.90$ months, range = 60 - 119 months, 8 participants in each age group, 20 female). Children were drawn from the same population and tested in the same context as children from Study 4A. Twelve additional children were excluded either because they failed to identify the rational and irrational third parties during both memory check questions (4 in the helping only condition, 5 in the punishment only condition) because they did not endorse sharing (1 in the helping only condition), or there was an experimental error (1 in the helping only condition, 1 in the punishment only condition).

Experimental design and procedure. The measures used in Study 4B were identical to those in Study 4A. To address the two possible alternative explanations described above, I tested children in two between-subject conditions. In the *Helping-only condition*, children's preference for the rational helper was compared with that for the irrational helper. The irrational helper gives two candies to the unfair divider, while the rational helper gives two candies to the recipient. In the *Punishment-only condition*, the irrational punisher who takes two candies away from the recipient was compared with the rational punisher who takes two candies away from the unfair divider (see Table 1).

If children's preference for helpers in Study 4A was due to their preference for givers to takers, I would expect no difference between irrational and rational helpers in the helping only condition. By contrast, if children in Study 4A made a judgment in terms of fairness, children in the current study would prefer rational over irrational helpers. Additionally, if children based their preference on those who maximize a total number of resources, there should be no difference in their preference between rational and irrational third parties as both rational and irrational third-party intervention result in the same amount of candies in each condition. By contrast, if those in Study 4A made a judgment in terms of fairness, they would prefer rational to irrational third parties in both conditions.

Results

Evaluations in Likert scale ratings. In the helping only condition, children's evaluation of rational helpers ($M = 2.08$, $SD = 1.59$) differed significantly from neutral, one sample t-tests, $t(39) = 8.25$, $p < .001$, implying that they evaluated rational helpers positively. In contrast, the evaluation of irrational helpers ($M = 0.60$, $SD = 2.05$) did not differ from neutral; $t(39) = 1.85$, $p > .07$. In the punishment only condition, in which children were presented with rational and irrational punishers, their evaluation of the rational punisher ($M = 2.10$, $SD = 1.28$) differed significantly from neutral; $t(39) = 10.4$, $p < .001$. The ratings for the irrational punisher ($M = 0.70$, $SD = 1.98$) differed from neutral; $t(39) = 2.24$, $p = .03$.

I ran a full LMM on the Likert scale ratings with condition (helping only vs. punishment only), intervention type (rational vs. irrational), and age and interactions with the predictors as fixed effects and subject ID as a random effect. The results revealed that there were no interaction effects involving condition (all ps > .60). This shows that children's evaluations of third-party actors were similar for the helping only and punishment only conditions. Critically,

there was a significant interaction between intervention type and age, LRT, χ^2 (1) = 19.08, b = -0.06, SE = 0.01, p < .001, indicating that children's evaluations of rational vs. irrational actors changed with age (see Figure 14B).

To unpack the interaction effect between intervention type and age, I ran separate LMMs for rational and irrational characters. I found that children's evaluation of rational third parties did not change depending on their age, χ^2 (1) < 1, p > .62. Whereas, children's age significantly predicted their rating of irrational characters, χ^2 (1) = 21.96, b = -0.06, SE = 0.01, p < .001, suggesting that children evaluated irrational third parties negatively as they grow older.

To assess the age at which children's rating of irrational third parties differs from that of rational ones, I computed the age point when confidence intervals no longer overlapped with each other. I found that around 82 months of age, children's rating of irrational actors becomes lower than rational ones in both the helping only (rational M = 2.00, 95% CI [1.55, 2.45]; irrational M = 1.04, 95% CI [0.59, 1.49]) and punishment only conditions (rational M = 2.04, 95% CI [1.60, 2.49]; irrational M = 1.08, 95% CI [0.63, 1.53]). These results suggest that children's understanding of the justifiability of third-party interventions becomes robust by 7 years of age. Therefore, the preference for helpers over punishers by children 7 years and older found in Study 4A cannot be explained by a mere preference for givers or for resource maximizers.

Social preference in forced-choice questions. In the helping-only condition, children's preference for rational helpers over irrational helpers increased with age from 50% in 5-year-olds to 88% in 9-year-olds when asked who they liked better. In friend and party questions, they showed a similar age-dependent increase in the preference (from 38% to 75% in the friend question and from 63% to 75% in the party question). In the punishment-only condition, similar

patterns were observed. Children's preference for rational punishers over irrational punishers increased with age from 50% in 5-year-olds to 88% in 9-year-olds when asked who they liked better. The preference for rational punishers increased with age from 25% to 88% in the friend question and from 25% to 63% in the party question.

Children's responses to the three forced-choice questions (rational third-party = 1, irrational third-party = 0) were averaged into a social preference score because their responses in the three forced-choice questions (friend, party and like better) were highly intercorrelated (Cronbach's $\alpha = .78$). A higher score indicates a stronger preference for rational over irrational third parties. The results from a full GLM including condition (helping only vs. punishment only), age and the interaction between condition and age as fixed effects revealed that there was neither main effect nor interaction effect involving condition (all *p*s > .25). However, I found a significant effect of age on children's social preference score, $\chi^2 (1) = 9.06$, $b = 0.01$, $SE = 0.004$, $p < .01$ (see Figure 14D). This suggests that with increasing age, children were more likely to prefer rational over irrational actors.

To assess the age at which children's preference for rational over irrational third parties becomes reliable, I computed the age point when confidence intervals no longer overlapped with the value of 1.5, which is the social preference score expected by chance. Results revealed that from 85 months of age, confidence intervals no longer overlapped with 1.5 ($M = 1.79$, 95% CI [1.502, 2.14]). This suggests that a preference for rational over irrational third parties becomes more reliable around age 7. Taken together, the findings from social preference scores as well as those from Likert scale ratings support that children gain a better understanding of proper targets of third-party interventions around 7 years of age.

Trait attribution. In the helping-only condition, children attributed empathy (80%; binomial test, $p < .001$) and generosity (75%; $p < .01$) to rational helpers than to irrational helpers. This suggests that even though both helpers showed the same giving behaviors, children infer warmth-related traits from rational helpers who restored recipients' loss rather than from irrational ones who rewarded selfish dividers. Also, children in the helping only condition attributed aggression to irrational helpers (80%; $p < .001$). However, their attribution of norm enforcement was not significant ($p > .26$; see Table 2). It is perhaps not surprising that children had difficulties with attributing norm enforcement to one of the helpers because the focus of helping is to restore recipients' loss, not to enforce norms on perpetrators.

In the punishment-only condition, I predicted that children would attribute generosity, empathy and norm enforcement to rational punishers while attributing aggression to irrational punishers. However, this was not the case: there were no consistent attributions of traits to either kind of punisher (all ps > .08; see Table 2). One possibility is that the taking behavior that both punishers displayed might have made children perceive both punishers as an aggressive individual regardless of their rationality, preventing them from appreciating underlying personality traits of justifiable vs. unjustifiable punishment. These results in the punishment-only condition contrast with children's successful trait attribution in the helping-only condition in which both helpers did not show taking behavior or Study 4A in which only one of two third parties showed taking behavior.

Discussion of Study 4B

Study 4B revealed that, from age 7, children like rational actors who decrease inequality more than irrational actors who increase inequality in both continuous Likert scale ratings and

forced-choice measures. This shows that by 7 years of age, children have a sophisticated understanding of justifiable third-party intervention.

I note that there was no significant effect of condition for children's rating of rational helpers and rational punishers. However, in contrast to Study 4A which aimed at directly comparing helpers and punishers, Study 4B with its focus on rational versus irrational intervention asked children to evaluate helpers and punishers as a between subject variable. As a consequence, children did not see helping and punishment side by side, which might explain why there was no significant difference in their evaluations of helpers and punishers– as long as helping and punishment reduced inequality.

Results of Study 4B address outstanding questions from Study 4A by showing that children's preference cannot be explained by a mere preference for givers over takers. This preference should have resulted in a non-significant difference in children's liking between rational and irrational helpers. This alternative hypothesis is refuted by the current findings that children attended to whether helping decreased or increased inequality. This study also rules out that children preferred helpers to punishers because helping maximizes the total number of resources. This alternative account predicted that there should be no difference in preference between rational and irrational third parties. This is refuted by children preferring rational over irrational third parties in both the helping-only and the punishment-only conditions. Also, here, children attended to the impact of third-party's action on fairness when evaluating them.

However, one could argue that it is still unclear how children think about third-party interventions because Study 4A and 4B always asked children to evaluate third-party actors, not the act of intervention per se. Specifically, it is possible that children might evaluate actors and actions differently. One hypothesis is that children would evaluate punishing action as negative

or neutral. This result would suggest that, although children rate punishers positively, they do not view punishment as a preferable way to intervene against unfairness perhaps due to the observed aggression (i.e., taking behavior) in punishment.

Alternatively, children would like punishing action more than helping action. That is, what children think is a proper way to intervene can differ from who they want to associate with. For example, even if they like helpers better than punishers as their friend, it is still possible they endorse punishment more than helping to uphold cooperative group norms and prevent future transgressions.

The purpose of Study 4C was (1) to replicate children's preference for helpers over punishers and (2) to examine whether children evaluate actions differently from the agents performing these actions or their evaluations of the actions correspond to their evaluation of the people performing them.

Study 4C: How Do Children Evaluate Helping and Punishing Actions?

Method

Participants. Our final sample were $N = 80$ 5- to 9-year-old children ($M = 89.17$ months, range = 59 - 118 months, $n = 16$ participants in each age group, 40 female). Children were recruited and tested at a museum in a Midwest town in the US. Demographic information such as race, education, and income were not asked. Ten additional children were excluded because they failed to identify the helper and punisher during both memory check questions (5), they did not endorse sharing (3) or there was interference from a parent or a friend (2).

Experimental design and procedure. Children heard the identical story with a third-party punisher and a helper as in Study 4A. The difference from Study 4A was that I included two new measures to examine children's evaluations of each third-party's action: (1) children's

agreement with punishing and helping actions on the Likert scale, ranging from -3 (totally disagree) to +3 (totally agree) and (2) their forced-choice preference between the two actions (punishing vs. helping). In Study 4C, I excluded trait attribution measures and two of the three forced-choice questions used in Study 4A and 4B (i.e., friend, party) to focus our research question on evaluations of actions. However, I still kept our primary measures (i.e., Likert scale ratings about helpers and punishers and who they "like better" between the two actors) in Study 4C to replicate children's preference for helpers over punishers.

After seeing the divider's unfair allocation, children were asked to rate their agreement with helping and punishing actions on the 7-point smiley face Likert scale. For example, I asked "Some children say that [the third-party actor's name] should take two candies away from [the unfair divider's name]. What do you think about the idea?". Subsequently, the same child was asked "Some children say that [the third-party actor's name] should give two candies to [the recipient's name]. What do you think about the idea?".

Furthermore, children were forced to choose what the third-party actor should do between punishment of the unfair divider and helping of the recipient ("What do you think [the third-party actor's name] should do? Do you think he should take two candies away from [the unfair divider's name] or give two candies to [the recipient's name]?"). Importantly, children were asked to indicate their evaluations about punishing and helping actions *before* they heard about the actual intervention decision that the third-party actor made.

I counterbalanced the order of the questions (ask punishing first vs. ask helping first), the order of agreement scale (agreement first vs. disagreement first) and the order of practice trials (practice positive statement first vs. practice negative statement first).

Results

Evaluations in Likert scale ratings. In terms of third-party actors, children's evaluations of helpers ($M = 2.10$, $SD = 1.21$) and punishers ($M = 1.09$, $SD = 1.77$) differed significantly from neutral (one sample t-tests, $t(79) = 15.55$, $p < .001$ and $t(79) = 5.49$, $p < .001$, respectively), suggesting that both the helper and punisher were rated positively. These results replicated the findings from Study 4A.

In terms of third-party actions, children's evaluations of helping ($M = 0.86$, $SD = 1.95$) differed significantly from neutral, suggesting that they viewed helping action positively, $t(79) = 3.95$, $p < .001$, whereas punishment ($M = -0.59$, $SD = 2.06$) differed significantly from neutral in the opposite direction, suggesting that children viewed punishing action negatively; $t(79) = -2.55$, $p < .05$.

The results from a full LMM on the Likert scale ratings with the target (actor vs. action), intervention type (helping vs. punishment), age and interactions among the predictors as fixed effects and subject ID as a random effect revealed that there is a significant three-way interaction effect involving target, intervention type and age (LRT, $\chi^2(1) = 5.32$, $p < .05$). To better understand the three-way interaction effect, I ran a separate LMM depending on the target (actor vs. action).

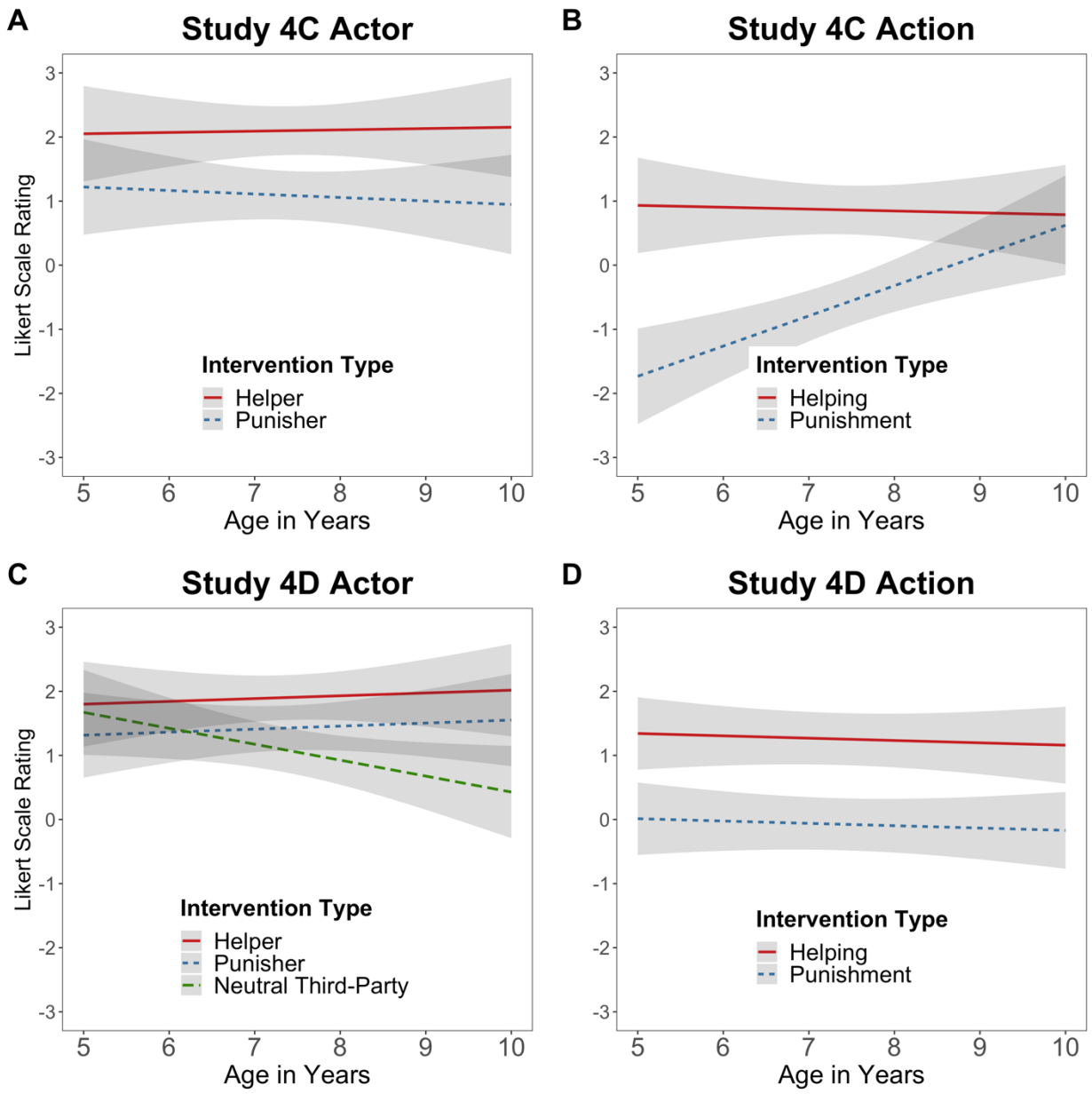

Figure 15. Likert scale ratings in Study 4C and 4D.

(A) Likert scale ratings of helpers and punishers in Study 4C. (B) Likert scale ratings of helping and punishing actions in Study 4C. (C) Likert scale ratings of helpers, punishers and neutral third parties in Study 4D. (B) Likert scale ratings of helping and punishing actions in Study 4D.

When it comes to actors, there was a significant main effect of intervention type (LRT, χ^2 (1) = 22.78, b = -1.01, SE = 0.20, p < .001), suggesting that children liked helpers more than

punishers (see Figure 15A). The findings show that children liked helpers more than punishers regardless of their age.

In terms of children's evaluations of actions, I found a significant interaction between intervention type and age (LRT, χ^2 (1) = 5.54, $p < .05$; see Figure 15B). To unpack this two-way interaction, I ran separate linear models for helping action and punishing action. For evaluations of the act of helping, there was a non-significant effect of age, $F(1, 78) < 1$, $p > .84$, indicating that children rated helping positively regardless of their age. However, regarding the act of punishment, I found a significant effect of age, $F(1, 78) = 9.68$, $b = 0.04$, $SE = 0.01$, $p < .01$. That is, with increasing age, children tend to view punishing action more positively. Further inspections of confidence intervals revealed that from 102 months of age (8.5 years), the confidence interval of punishing action ratings started to overlap with that of helping action ratings (helping $M = 0.83$, 95% CI [0.36, 1.30]; punishment $M = -0.08$, 95% CI [-0.56, 0.39]). This finding tentatively suggests that a tendency to view punishment as a proper way to intervene against unfairness becomes reliable between ages 8 and 9 (but see our follow-up Study 4D results for a discussion of the robustness of this effect).

Social preference in forced-choice questions. When asked who they liked better, a majority (71%) of children chose helpers over punishers (binomial test, 95% CI [0.60, 0.81], $p < .001$), replicating the results from Study 4A. When asked to choose between two third-party actions, a majority (63%) of children chose the helping action over the punishing action (binomial test, 95% CI [0.51, 0.73], $p < .05$), reporting that the third-party actor should help the recipient rather than to punish the unfair divider. The results from a GLM indicated that children preferred helping over punishment irrespective of the target (actor vs. action) or their age.

Discussion of Study 4C

In Study 4C, I replicated a preference for helpers over punishers in children. Both Likert-scale and forced-choice measures suggest that children like third-party helpers better than punishers irrespective of their age.

A new feature of Study 4C was that it assessed children's evaluations about third-party actions. I found that 5- to 9-year-olds evaluated helping action positively regardless of their age. By contrast, children's evaluation of punishing action became more positive between 8 and 9 years of age. However, I did not find a similar age-related trend in the forced-choice question, in which children had to choose which action is a better way to intervene. Here, most children chose helping over punishing action regardless of their age.

Overall, our findings suggest that across all age groups, children prefer helping over punishment. Even if children older than 8 start to view punishment more positively than younger children do, they still prefer helping over the punishing action when they must choose only one type of intervention.

In Study 4A and 4C, a major finding was that children view punishers positively, but they do not prefer punishers as much as they prefer helpers. However, it is still unclear whether children genuinely like third-party punishers. One possibility is that children rated punishers positively because of their general positivity bias towards any human characters. In other words, children's positive ratings of punishers might be inflated because of their mere liking of human characters in general, not because of the punishment the character performed.

Study 4D was designed to address this possibility by introducing a neutral third-party. If children's positive evaluations of punishers reflect their genuine liking, I would expect that children will like punishers more than the neutral third-party character. This result would suggest

that punishers gain additional reputational benefits by enacting punishment. Alternatively, if their positive evaluations of punishers are a mere positivity bias towards human characters irrespective of their actions, children should like the neutral third-party as much as the punisher. Such an outcome would suggest that punishers do not gain any additional reputations by enacting punishment.

The purpose of Study 4D was to replicate Study 4C and to assess if punishers gain additional positive evaluations compared to neutral third parties.

Study 4D: Do Children Evaluate Punishers More Positively Than Neutral Third Parties?

Method

Participants. Our final sample were $N = 82$ 5- to 9-year-old children ($M = 88.30$ months, range = 60 – 118 months, $n = 17$ 5-year-olds, $n = 17$ 6-year-olds, $n = 17$ 7-year-olds, $n = 17$ 8-year-olds, $n = 14$ 9-year-olds, 42 female). Children were drawn from the same population and tested in the same context as children from Study 4C. Five additional children were excluded because they failed to identify the third-party actors during both memory check questions.

Experimental design and procedure. The procedure was identical to that of Study 4C except that children were introduced to a neutral third-party. At the beginning of the story, the experimenter said that the neutral third-party character is busy and cannot join the candy game that the other four characters—a divider, a recipient, a helper, and a punisher — are involved with. Then, the neutral third party left and did not appear during the rest of the story. That is, the neutral third party neither watched any of the events nor got involved in any way. The neutral third-party character was presented to children again along with the other four characters when the experimenter asked the test questions. Therefore, it appeared only twice during the entire testing session: At the beginning and the end (test phase) of the study.

All the measures were identical to those used in Study 4C except that I added (1) the Likert scale rating of the neutral third-party and (2) two additional forced-choice preference questions involving the neutral third-party (helper vs. neutral, punisher vs. neutral).

Results

Evaluations in Likert scale ratings. Children's evaluations of helpers (M = 1.90, SD = 1.60), punishers (M = 1.43, SD = 1.70) and neutral third parties (M = 1.09, SD = 1.62) differed significantly from the neutral value of zero (one sample t-tests, $t(81) = 10.77$, $p < .001$; $t(81) = 7.60$, $p < .001$; $t(81) = 6.07$, $p < .001$, respectively), suggesting that all three third-party actors were viewed positively (see Figure 15C).

The results from an LMM with age, intervention type and interactions with the predictors as fixed effects and subject ID as a random effect revealed that there was a significant effect of intervention type (LRT, $\chi^2(2) = 11.46$, $p < .01$). The results indicated that children liked helpers more than punishers (b = -0.48, SE = 0.24, $p < .05$) and liked helpers more than neutral third parties (b = -0.82, SE = 0.24, $p < .001$). However, there was no difference in evaluations between punishers and neutral third parties (b = -0.34, SE = 0.24, $p > .15$).

When it comes to actions, children's evaluations of helping (M = 1.26, SD = 1.97) differed significantly from zero, suggesting that they viewed helping action positively (one sample t-test, $t(81) = 5.76$, $p < .001$). This result replicated the findings from Study 4C. By contrast, children's ratings of punishment (M = -0.07, SD = 2.20) did not differ from zero, suggesting that children viewed the punishing action neutrally (one sample t-test, $t(81) = -0.30$, $p > .76$).

Next, to test the effect of target (actor vs. action) on the ratings, I ran an LMM on the Likert scale ratings with the target (actor vs. action), intervention type (helping vs. punishment),

age and interactions with the predictors as fixed effects and subject ID as a random effect. There was a significant interaction between the target and intervention type (LRT, χ^2 (1) = 4.99, p < .05; see Figure 15D).

To unpack the two-way interaction between the target (actor vs. action) and intervention type (helping vs. punishment), I ran separate LMMs for the actor and action, respectively. Children liked helpers more than punishers (χ^2 (1) = 4.61, b = -0.48, SE = 0.22, p < .05). Also, they liked helping action more than punishing action (χ^2 (1) = 22.45, b = -1.33, SE = 0.26, p < .001). That is, regardless of target (actor vs. action), children liked helping more than punishment. The interaction effect between the target and intervention type was because children's preference for helping over punishment is more pronounced in their ratings of actions than those of actors.

Social preference in forced-choice questions. When asked who they liked better, children's preference for helpers over neutral third parties increased with age (from 41% in 5-year-olds to 100% in 9-year-olds). Similarly, the preference for punishers over neutral third parties increased from 29% to 86%, and the preference for helpers over punishers increased from 35% to 93%.

To examine the effect of age on children's forced-choice preferences, I ran three separate GLMs on children's choice in (1) helper vs. neutral third-party, (2) punisher vs. neutral third-party, and (3) helper vs. punisher. First, the results from a GLM with a binary response term (helper = 1, neutral third-party = 0) indicated that age predicted children's preference (LRT, χ^2 (1) = 11.04, b = 0.05, SE = 0.02, p < .001). That is, children are more likely to prefer helpers over neutral third parties as they grow older (see Figure 16A). Closer inspections of confidence

intervals revealed that confidence intervals no longer overlap with chance from 79 months (6.6 years) (M = 0.64, 95% CI [0.51, 0.75]).

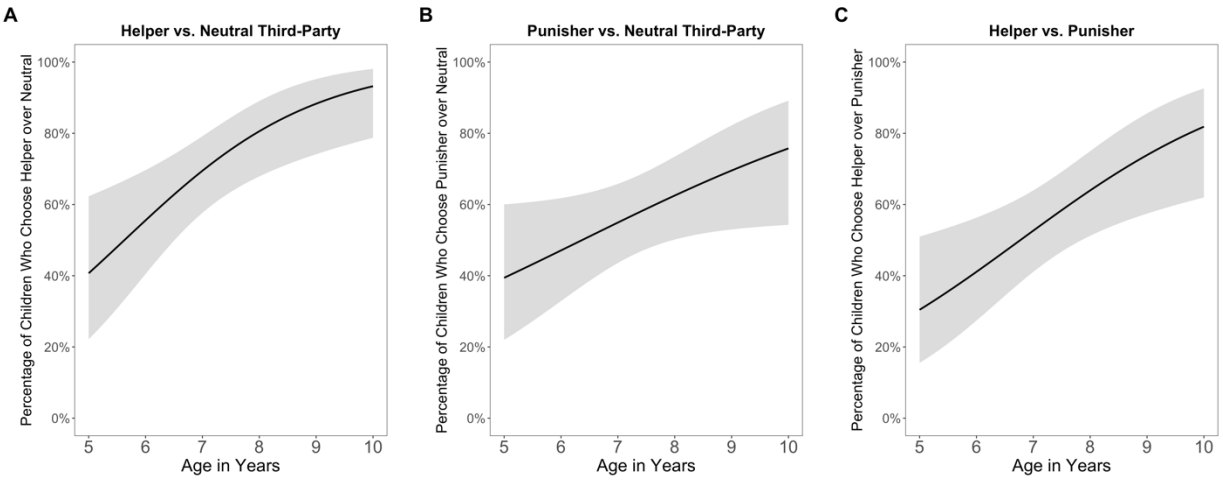

Figure 16. Forced-choice social preferences in Study 4D.

(A) Higher percentage indicates children's preference for helpers over neutral third parties. (B) Higher percentage indicates children's preference for punishers over neutral third parties. (C) Higher percentage indicates children's preference for helpers over punishers.

Second, the results from a GLM with a binary response term (punisher = 1, neutral third-party = 0) revealed that age predicted children's preference (LRT, χ^2 (1) = 4.06, b = 0.03, SE = 0.01, $p < .05$); Figure 16B), suggesting that children were more likely to prefer punishers over neutral third parties with age. Confidence intervals no longer overlap with chance from 96 months (8 years) of age (M = 0.63, 95% CI [0.502, 0.73]).

Third, the results from the GLM with a binary response term (helper = 1, punisher = 0) revealed that age significantly predicted children's preference (LRT, χ^2 (1) = 8.57, b = 0.04, SE = 0.01, $p < .01$; Figure 16C), suggesting that their preference for helpers to punishers increased with their age. From 95 months of age (7.9 years), confidence intervals no longer included chance (M = 0.63, 95% CI [0.51, 0.74]).

When asked to choose between two actions, most children (70%) chose helping over the punishing action (binomial test, 95% CI [0.58, 0.79], $p < .001$), replicating the findings from Study 4C. The results from a GLMM indicated that there was no effect of age on children's choice between helping and punishing action, suggesting that children prefer helping over punishment regardless of age.

Discussion of Study 4D

Study 4D replicated the findings from Study 4C that children liked helpers more than punishers on the Likert scale ratings and that they preferred helping action over punishing action. However, unlike in Study 4C in which children rated punishment more positively with age, I did not find the same age effect in Study 4D. That is, regardless of their age, children rated punishment neutrally on the Likert scale. I conclude that the age effect found in Study 4C was not reliable enough because of the small sample size in each age group or the narrow age range.

I found that children liked helpers more than neutral third parties. Interestingly, children's ratings of punishers and neutral third parties did not differ significantly from each other, suggesting that punishers do not receive any additional positive evaluations from children compared to neutral third parties. In the forced-choice task, however, children's tendency to prefer punishers over neutral third parties becomes stronger around age 8. These results suggest that the enactment of punishment does not entail additional reputational benefits in children younger than 8.

In contrast to Study 4A and 4C, I did not replicate children's preference for helpers over punishers in the forced-choice question. In Study 4D, children's preference for both third parties did not differ significantly, and their preference for helpers to punishers becomes stronger around age 8. I speculate that young children might have had difficulties with comparing three pairs of

characters (Note that I found an age effect in all three forced-choice tasks; see Figure 16). The confusion might have been more pronounced in young children when they had to compare three different characters with each other in the forced-choice tasks than when they had to report their liking for one character at a time on the Likert scale in which I did not find an age effect.

Discussion of Studies 4A-4D

This series of studies provides new evidence on children's evaluations of third-party interventions against fairness norm violations. This was possible, in part, because of our novel comparison between children's evaluations of those who punish selfish resource dividers with those who help recipients of selfish resource allocations. I started with the recent evidence showing that adults generally view helpers more favorably compared to punishers (Jordan et al., 2016; Patil et al., 2018; Raihani & Bshary, 2015a). For a first look at this phenomenon in children, I tested two competing hypotheses: One hypothesis assumed developmental continuity, proposing that this preference might be a deep-seated psychological phenomenon that is present already in children. The counterhypothesis conjectured that there would be a developmental discontinuity, with young children focusing more on superficial features of the actions, judging helpers as positive and the more aggressive punishers as negative, while older children acknowledge potential benefits of punishment and therefore evaluating punishers positively, and perhaps as positively as helpers.

Overall, the current findings are consistent with the first hypothesis that children would share a similar preference with adults. Across studies, I found that children between 5 and 9 years of age evaluated punishers positively. However, they liked helpers more than punishers. For example, children rated helpers more positively than punishers on the Likert scale (Study 4C and 4D). Similarly, they chose helpers over punishers in the forced-choice measures (Study 4A

and 4C). Furthermore, in Study 4A, I found that children inferred warmth-related traits (generosity, empathy) from helpers, while they inferred conflict-related traits (aggression, norm enforcement) from punishers, showing their reasoning behind the preference for helpers.

I also ruled out two alternative hypotheses according to which children's preference for helpers could be due to their preference for givers or for resource maximizers, respectively. Contrary to these alternative hypotheses, in Study 4B, I found that children aged 7 years and older liked rational third parties who decreased inequality more than irrational ones who increased inequality. This suggests that their preference for helpers can neither be explained by a mere preference for givers over takers nor by a preference for resource maximizers. In fact, from 7 years on, children conceived of third-party interventions as aimed at identifying a proper target to reduce inequality rather than as indiscriminate giving or resource maximizing.

Furthermore, I found that children's preference for helping over punishment is not just limited to judgments of third-party actors. When I assessed children's preference for the action per se, they liked helping actions more than punishing actions. Helping actions were always rated positively while punishing actions were rated negatively or neutrally (Study 4C and 4D). I found similar results in the forced-choice task: Children chose helping over punishment irrespective of their age (Study 4C and 4D).

The current study contributes to the literature by moving beyond the focus on punishment alone and probing children's thinking about punishment and helping side by side. Prior developmental research focused on comparing punishers with third parties such as onlookers who choose not to intervene after witnessing a transgression (e.g., Vaish et al., 2016) or givers who reward a transgressor (e.g., Hamlin et al., 2011), which might have led to inflating children's preference for punishers. Instead, the current study compared punishment with

helping, a valid and common form of third-party intervention. Additionally, our study assessed children's evaluations of punishment intervention per se and revealed a subtle but meaningful difference in understanding punishers vs. punishment, which was especially remarkable in young children. With the use of various measures and comparisons, the current study provided a more comprehensive understanding of the development of third-party punishment in children.

Limitations and Future Directions

One potential concern is that the developmental pattern found in the current experiments is due to cognitive demands inherent in our tasks, rather than reflecting changes in fairness judgments per se. For example, younger children might have had difficulty keeping track of the identities and behaviors of each actor in our story. However, this is unlikely for two reasons. First, I included two memory checks, revealing that most children understood the events and identified actors in the story correctly (90% in Study 4A and 4B, 94% in Study 4C and 4D), while the few who failed both memory checks were excluded from our analyses. Second, our data do not support the idea that young children found the task too difficult. If young children had problems with understanding the story, they should have shown results different from those of older children across a wide range of measures. However, in the majority of our measures, I did not find significant age differences. For example, regardless of their age, children evaluated both helpers and punishers positively and preferred helpers over punishers. Additionally, there were no age differences in children's ability to attribute relevant traits to a third-party (except for attribution of norm enforcement in Study 4A). Hence, it is unlikely that the task difficulty affected a few specific measures selectively. In summary, I believe that the age differences found in our measures reflect the development of children's understanding of fairness and third-party interventions rather than the cognitive demands of our tasks.

The current study elucidates the development of children's evaluations of third-party punishment and helping in the context of fairness norm violations. One critical question to consider is *why* children and adults show this preference. A possible explanation is that people prefer those who show an empathetic concern for the well-being of others. People might consider helpers as someone who is trustworthy and dependable when in need, and thus want to associate with helpers rather than with punishers (Jordan et al., 2016; Patil, Dhaliwal, & Cushman, 2018). Another reason could be that helping is regarded as an empathetic and generous act, whereas punishment, even though justified, is regarded as aggressive. As a consequence, people may acknowledge that punishers are a necessary asset to uphold group norms but may be less likely to choose them as social partners due to the perceived aggression or dominance they display when enforcing norms (Gordon, Madden, & Lea, 2014). Another possibility is that children prefer those who establish equality and maximize resources simultaneously. One potential reason for children preferring helpers over punishers is that helpers not only establish equality but also maximize resources (enlarge the pie), whereas punishers establish equality but minimize available resources. Not surprisingly, adults seem to have a similar preference: they prefer an option that not only establishes equality but enlarge the pie at the same time over a punitive intervention (FeldmanHall et al., 2014; Heffner & FeldmanHall, 2019). Our study suggests that children do not merely like anyone who enlarges the pie (e.g., when helping increases inequality). However, it is still possible and indeed likely that children endorse resource maximization per se, as long as it does not increase inequality. Future research should investigate reasons underlying the preference for helping over punishment in children.

Another critical question is whether children's preference for helpers found in fairness norm violations in the current study can be generalized to other moral contexts. For example, in

the context of physical harm, children might prefer punishment over helping because the physical harm inflicted on the victim might be considered as irrecoverable and thus not compensable. On the other hand, physical harm also often triggers an urge to comfort a victim or relieve pain, so children might regard punishment of perpetrators without a remedy for the victim as cruel. Similar arguments can be made about property damage or theft, opening many possibilities for future research across diverse moral domains.

Chapter 6: General Discussion

The goal of my dissertation was to investigate developmental trajectories and motivations underlying third-party punishment in children. The present work provides insight into three key issues: (1) when and how third-party punishment develops in children, (2) what motivates children to punish, and (3) how their evaluation of punishment compares to another form of intervention: third-party helping. Across a series of experiments, I have found that third-party punishment in children is driven by a concern for fairness. Concretely, when children can decide the degree of punishment, with increasing age, they fine-tune the exact amount of punishment needed to restore equality between two other people (Study 1). Similarly, as they grow older, children become more likely to pay a cost in order to stop the selfish divider from having all coins (Study 2). Findings from these two studies suggest that children are motivated not only to prevent inequality but also to actively create equality between individuals. Further, in Study 4 in which children heard about third-party interventions, with age, children prefer a third-party who decreased inequality over a third-party who increased inequality. Together, this provides converging evidence that over development, children think of third-party interventions as a way to reduce unfairness between two other people.

Furthermore, results from Study 2 and 3 suggest that children's third-party punishment is robust against self-focused factors. For example, their rate of third-party punishment was not easily affected by their personal experience of (un)fairness (Study 2) or the possibility of encountering the same social partner in the future (Study 3). Therefore, it seems that, in a certain

context, children do not use third-party punishment in a strategic manner to enhance or protect their personal interest.

Although the studies described above demonstrated and highlighted the importance of third-party punishment in enforcing fairness norms, findings from Study 4 questioned the focus on punishment as a sole intervention against unfairness. To elaborate, Study 4 found that children assign positive evaluations to a third-party punisher, but they tend to prefer a third-party who compensated a disadvantaged recipient over a third-party punisher. These findings suggest that the role of punishment might have been overrated. Existing literature with adults and children focused exclusively on the role of punishment in cooperation. However, these studies presented punishment as the only way to intervene against a moral transgression without allowing an alternative intervention option (e.g., Fehr & Fischbacher, 2004a; Fehr & Gächter, 2002; House et al., 2020; McAuliffe et al., 2015). Hence, Study 4 casts doubt on the degree to which punishment is endorsed in real life and calls for research on alternative interventions other than punishment.

This dissertation investigated the development of third-party punishment and its underlying motives. The findings across the studies open up the possibility of three exciting future directions. First, more work is needed to better understand how children's use of punishment compares to other types of third-party intervention. As mentioned above, it is possible that children in the current studies enacted costly punishment because it was the only option available to intervene in the situation. However, if children had had an alternative option (e.g., compensation or redistribution), they might have enacted punishment less often. Although the current studies primarily explored the development of monetary punishment, it is not the only way to enforce social norms. In fact, there are numerous other ways to respond to a moral

transgression in real life. These alternative interventions include tattling, reproaching verbally or forgiving. Future research should examine when and how children use these alternative interventions as a way to respond to a transgression.

Second, future research should examine the extent to which children use third-party interventions strategically. In Study 3, I found that children do not use punishment differentially when they were told to interact with the same divider vs. a different divider, suggesting that they do not use punishment as a way to manipulate their future social partners. However, it is still unclear whether children's lack of strategic interventions is limited to a punishment context or not. To elaborate, it is possible that children are not strategic only when they have to intervene by punishing, but they could still show strategic interventions when the context does not involve punishment. One way to address this question could be to examine if children use another behavior (e.g., rewarding of a fair allocation) strategically when they are told to interact with the same divider (vs. a different divider). By comparing children's use of reward with their use of punishment, I would be able to determine whether children's insensitivity to their future partner is limited to the punishment context specifically or whether the insensitivity could be generalized to another context as well.

Lastly, future research should also investigate how children from different populations and cultures reason about and engage in third-party punishment. While I selected children from the US based upon prior studies, there are cross-cultural variations in fairness norms and behaviors across both adults and children (Blake et al., 2015; Herrmann, Thöni, & Gächter, 2008; Henrich et al., 2006; House et al., 2013). To be specific, most societies show punishment of selfish behaviors (Henrich et al., 2006; Hermann et al., 2008), showing that punishment of selfishness may be universal. Whereas, when it comes to punishment of cooperative behavior,

there are great variabilities across societies. Participants from collectivistic societies (e.g., Muscat, Athens) punish those who cooperated equally or more than themselves, leading to less cooperation in the group (Hermann et al., 2008). This suggests that punishment is not always directed towards antisocial behaviors. Instead, it could be used as a way to define social norms even if those antisocial norms could potentially undermine cooperation in the society. Therefore, it is a critical task for future research to assess similarities and differences in the developmental trajectory of fairness across different populations.

As a part of this plan, I will examine how children's evaluations about hyper-generous sharing (e.g., keeping 0 for the self and giving all resources to a partner) differs between US and Korea. One prediction is that children from Korea would punish hyper-generous sharing more often than those from US. Children from collectivistic culture could show more punishment towards hyper-generous sharing because most interactions in this culture could be limited to close-knit, ingroup-oriented relationships, and thus be suspicious of strangers who seem too generous without a clear reason (Hermann et al., 2008). I hope that this future research could address how culture shapes social norms and its punishment in early childhood.

Bibliography

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., … Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, *1*, 357–366.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*, 596–612.

Atance, C. M., & Meltzoff, A. N. (2005). My future self: Young children's ability to anticipate and explain future states. *Cognitive Development*, *20*, 341–361.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*, 59–78.

Bernhard, R., Martin, J., & Warneken, F. (in press). Why do children punish? Fair outcomes matter more than intent in children's second- and third-party punishment. *Journal of Experimental Child Psychology.*

Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair" Eight-year-olds reject two forms of inequity. *Cognition*, *120*, 215–224.

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., Kleutsch, L., Kramer, K. L., Ross, E., Vongsachang, H., Wrangham, R., & Warneken, F. (2015). The ontogeny of fairness in seven societies. *Nature*, *528*, 258–261.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 3531–3535.

Bregant, J., Shaw, A., & Kinzler, K. D. (2016). Intuitive jurisprudence: Early reasoning about the functions of punishment. *Journal of Empirical Legal Studies*, *13*, 693–717.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*, 378-400.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*, 1-28.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, *39*, 268–277.

Cooley, S., & Killen, M. (2015). Children's evaluations of resource allocation in the context of group norms. *Developmental Psychology*, *51*, 554–563.

Craig, C. D., & Sprang, G. (2007). Trauma exposure and child abuse potential: Investigating the cycle of violence. *The American journal of orthopsychiatry*, *77*, 296–305.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*, 794–796.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, *38*, 734–743.

Deschamps, T. D., Eason, A. E., & Sommerville, J. A. (2015). Infants associate praise and admonishment with fair and unfair individuals. *Infancy*, 1–27.

Dodge, K. A., Bates, J. E., & Pettit, G. S. (1990). In the cycle of violence through what intrapersonal mechanisms. *Science*, *250*, 1678–1683.

Elenbaas, L., Rizzo, M. T., Cooley, S., & Killen, M. (2016). Rectifying social inequalities in a resource allocation task. *Cognition*, *155*, 176–187.

Engelmann, J. M., & Rapp, D. J. (2018). The influence of reputational concerns on children's prosociality. *Current Opinion in Psychology*, *20*, 92–95.

Fagan, A. A. (2005). The relationship between adolescent physical abuse and criminal offending: Support for an enduring and generalized cycle of violence. *Journal of Family Violence*, *20*, 279–290.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*, 1079–1083.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*, 785–791.

Fehr, E., & Fischbacher, U. (2004a). Third party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.

Fehr, E., & Fischbacher, U. (2004b). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*, 185–190.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*, 1–25.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Fehr, E., Hoff, K., & Kshetramade, M. (2008). Spite and development. *The American Economic Review*, *98*, 494–499.

Fehr, E., & Schmidt, K. M. (1999). A Theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*, 817–868.

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Communications*, *5*, 5306.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*, 1360–1383.

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental Science*, *14*, 1012–1020.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language*, *1*, 158–171.

Gordon, D. S., Madden, J. R., & Lea, S. E. G. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PLoS ONE*, *9*, 1–10.

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition*, *133*, 97–103.

Gurerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*, 108–111.

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 19931–6.

Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language*, *7*, 120–144.

Heffner, J., & FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. *Scientific Reports*, *9*, 1–13.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich., N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.

Hermann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*, 1362–1367.

House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., Hewlett, B.S., McElreath, R., & Laurence, S. (2013). Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences*, *110*, 14586–14591.

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., … Silk, J. B. (2020). Social norms and cultural diversity in the development of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences*, *287*, 20192794.

Jeffreys, H. (1961). *Theory of probability*. Oxford, England: Oxford University Press.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, *102*, 192–194.

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences*, *111*, 12710–12715.

Jordan, J., Hoffman, M., Bloom, P., & Rand, D. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*, 473–476.

Kanakogi, Y., Inoue, Y., Matsuda, G., Butler, D., Hiraki, K., & Myowa-Yamakoshi, M. (2017). Preverbal infants affirm third-party interventions that protect victims from aggressors. *Nature Human Behaviour*, *1*, 0037.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*, 1–14.

Liu, D., & Vanderbilt, K. E. (2013). Children learn from and about variability between people. In M. R. Banaji, & S. A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 197-200). New York, NY: Oxford University Press.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences*, *108*, 11375–11380.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1–10.

McAuliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The developmental foundations of human fairness. *Nature Human Behaviour*, *1*, 0042.

Mendes, N., Steinbeis, N., Bueno-Guerra, N., Call, J., & Singer, T. (2018). Preschool children and chimpanzees incur costs to watch punishment of antisocial others. *Nature Human Behaviour*, *2*, 45–51.

Meristo, M., & Surian, L. (2014). Infants distinguish antisocial actions directed towards fair and unfair agents. *PLoS ONE*, *9*, 1–7.

Nelissen, R. M. A. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*, 242–248.

Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making*, *4*, 543–553.

Patil, I., Dhaliwal, N., & Cushman, F. A. (2018, May 23). Reputational and cooperative benefits of third-party helping. doi:10.31234/osf.io/c3bsj

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary psychology and criminal justice: A recalibrational theory of punishment and reconciliation. In Høgh-Olesen, H. (Ed.), *Human Morality and Sociality Evolutionary and Comparative Perspectives* (pp. 72–131). New York: Palgrave MacMillan.

Pfattheicher, S., Sassenrath, C., & Keller, J. (2019). Compassion magnifies third-party punishment. *Journal of personality and social psychology*, *117*, 124–141.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2018). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-137, <URL: https://CRAN.R-project.org/package=nlme>.

Raihani, N. J., & Bshary, R. (2015a). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, *69*, 993–1003.

Raihani, N. J., & Bshary, R. (2015b). The reputation of punishers. *Trends in Ecology and Evolution*, *30*, 98–103.

Raihani, N. J., & Bshary, R. (2019). Punishment: one tool, many uses. *Evolutionary Human Sciences*, *1*, e12, 1–26.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rizzo, M. T., & Killen, M. (2016). Children's understanding of equity in the context of inequality. *The British Journal of Developmental Psychology*, *34*, 569–581.

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, *141*, 382–395.

Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science*, *23*, 196–204.

Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. *PLoS ONE*, *8*(3): e59510.

Smith, C. E., & Warneken, F. (2016). Children's reasoning about distributive and retributive justice across development. *Developmental Psychology*, *52*, 613– 628.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K. W., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. *Journal of Neuroscience*, *38*, 2944–2954.

Staub, E., & Vollhardt, J. (2008). Altruism born of suffering: the roots of caring and helping after victimization and other trauma. *The American journal of orthopsychiatry*, *78*, 267–280.

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *The British Journal of Developmental Psychology*, *29*, 124–130.

Vaish, A., Herrmann, E., Markmann, C., & Tomasello, M. (2016). Preschoolers value those who sanction non-cooperators. *Cognition*, *153*, 43–51.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. C. (2019). Rank-normalization, folding, and localization: An improved Rhat for assessing convergence of MCMC. arXiv preprint arXiv:1903.08008.

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., … Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 20364–20368.

Yamagishi, T., Li, Y., Fermin, A. S. R., Kanai, R., Takagishi, H., Matsumoto, Y., … Sakagami, M. (2017). Behavioural differences and neural substrates of altruistic and spiteful punishment. *Scientific Reports*, *7*, 1–8.

Yudkin, D. A., Van Bavel, J. J., & Rhodes, M. (2020). Young children police group members at personal cost. *Journal of Experimental Psychology: General, 149*, 182-191.